WMT 2025

Tenth Conference on Machine Translation

Proceedings of the Conference

November 8-9, 2025

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-341-8

Introduction

The Tenth Conference on Machine Translation (WMT 2025) took place on November 8–9, 2025, immediately following the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025) in Suzhou, China.

This is the tenth time WMT has been held as a conference. The first time WMT was held as a conference was at ACL 2016 in Berlin, Germany, the second time at EMNLP 2017 in Copenhagen, Denmark, the third time at EMNLP 2028 in Brussels, Belgium, the fourth time at ACL 2019 in Florence, Italy, the fifth time at EMNLP-2020, which was held as an online event due to the COVID-19 pandemic, the sixth time at EMNLP 2021 at Punta Cana, Dominican Republic, the seventh time at EMNLP 2022 in Abu Dhabi, United Arab Emirates, the eight time at EMNLP 2023 in Singapore, and the ninth time at EMNLP 2024 in Miami, USA. Prior to being a conference, WMT was held 10 times as a workshop. WMT was held for the first time at HLT-NAACL 2006 in New York City, USA. In the following years the Workshop on Statistical Machine Translation was held at ACL 2007 in Prague, Czech Republic, ACL 2008, Columbus, Ohio, USA, EACL 2009 in Athens, Greece, ACL 2010 in Uppsala, Sweden, EMNLP 2011 in Edinburgh, Scotland, NAACL 2012 in Montreal, Canada, ACL 2013 in Sofia, Bulgaria, ACL 2014 in Baltimore, USA, EMNLP 2015 in Lisbon, Portugal.

The focus of our conference is to bring together researchers from the area of machine translation and invite selected research papers to be presented at the conference.

Prior to the conference, in addition to soliciting relevant papers for review and possible presentation, we conducted 10 shared tasks. These consisted of 5 translation tasks: General Translation, Low-Resource Indic Language Translation, Terminology Translation, Creole Language Translation, and Model Compression, two evaluation tasks: MT Test Suites ("Help us break LLMs, Vol. 2") and Automated Translation Quality Evaluation Systems, two multilingual tasks: Multilingual Instruction and LLMs with Limited Resources for Slavic Languages and finally the Open Language Data Initiative.

The results of all shared tasks were announced at the conference, and these proceedings also include overview papers for the shared tasks, summarizing the results, as well as providing information about the data used and any procedures that were followed in conducting or scoring the tasks. In addition, there are short papers from each participating team that describe their underlying system in greater detail.

Like in previous years, we have received a far larger number of submissions than we could accept for presentation. WMT 2025 has received 60 full research paper submissions (not counting withdrawn submissions). In total, WMT 2025 featured 21 full research paper presentations and 79 shared task presentations.

The invited talk was given by Longyue Wang from Alibaba, China.

We would like to thank the members of the Program Committee for their timely reviews. We also would like to thank the participants of the shared task and all the other volunteers who helped with the evaluations.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz

Co-Chairs

Organizing Committee

Chairs

Barry Haddow, University of Edinburgh
Tom Kocmi, Cohere
Philipp Koehn, Johns Hopkins University
Christof Monz, University of Amsterdam

Program Committee

Chairs

Idris Abdulmumin, University of Pretoria
Laurie Burchell, Common Crawl Foundation
Isaac Caswell, Google Research
Daryna Dementieva, Technical University of Munich
Marion Di Marco, TUM
Lukas Edman, TU Munich
Mark Fishel, University of Tartu
Barry Haddow, University of Edinburgh and Aveni
Tom Kocmi, Cohere
Philipp Koehn, Johns Hopkins University
Jean Maillard, Meta AI
Christof Monz, University of Amsterdam
Shu Okabe, TUM Heilbronn
Lisa Yankovskaya, University of Tartu

Program Committee

Sadaf Abdul Rauf, Fatima Jinnah Women Unversity
Kolawole Adebayo, Dublin City University
David Ifeoluwa Adelani, McGill University / MILA
Aniefon Akpan, Univeristy of Uyo
Asim Alaskar, King Saud University
Priscilla Amuok, Masakhane
Antonios Anastasopoulos, George Mason University
Anietie Andy, Howard University
Ramakrishna Appicharla, Indian Institute of Technology Patna
Ali Araabi, University of Amsterdam
Anuoluwapo Aremu, Lelapa AI, University of Trento, Masakhane
Joseph Attieh, University of Helsinki
Mikko Aulamo, University of Helsinki
Eleftherios Avramidis, German Research Center for Artificial Intelligence (DFKI)
Olumide Awokoya, Masakhane
Seth Aycock, University of Amsterdam
Fatemeh Azadi, National Research Council Canada
Selene Baez Santamaria, University of Zurich
Parnia Bahar, Apple
Rachel Bawden, Inria
Meriem Beloucif, Uppsala University
Nathaniel Berger, Heidelberg University
Toms Bergmanis, Tilde
Siddharth Betala, Indian Institute of Technology Madras
Alexandra Birch, University of Edinburgh
Ondřej Bojar, Charles University, MFF UFAL
Maharaj Brahma, Indian Institute of Technology Hyderabad
Eleftheria Briakou, Google
José G. C. De Souza, Unbabel

Marine Carpuat, University of Maryland
Antonio Castaldo, University of Naples L'Orientale"
Sheila Castilho, Dublin City University
Rajen Chatterjee, Apple Inc.
Pinzhen Chen, University of Edinburgh and Aveni
Colin Cherry, Google
Pranjal Chitale, Indian Institute of Technology, Madras
Katsuki Chousa, NTT
Chenhui Chu, Kyoto University
Josep Crego, CHAPSVISION
David Dale, Meta AI
Brian Davis, Dublin City University
Ona De Gibert, University of Helsinki
Nisansa De Silva, University of Moratuwa
Hiroyuki Deguchi, NTT, Inc.
Steve Deneefe, RWS Language Weaver
Michael Denkowski, Amazon
Sourabh Deoghare, IIT Bombay
Daniel Deutsch, Google
Giorgio Maria Di Nunzio, University of Padua
Shuoyang Ding, NVIDIA
Miguel Domingo, Universitat Politècnica de València
Sören Dreano, Dublin City University
Kevin Duh, Johns Hopkins University
Koel Dutta Chowdhury, Saarland Informatics Campus,Saarland University
Hiroshi Echizen'ya, Hokkai-Gakuen University
Hafsteinn Einarsson, University of Iceland
Muhammad Elnokrashy, Microsoft
Grant Erdmann, Air Force Research Laboratory
Carlos Escolano, Universitat Politècnica de Catalunya, Barcelona Supercomputing Center
Miquel Esplà-Gomis, Universitat d'Alacant
Marcello Federico, Amazon
Patrick Fernandes, Carnegie Mellon University, Instituto de Telecomunicações
Aloka Fernando, University of Moratuwa
Edoardo Ferrante, Conseggio pe-o patrimonio linguistico ligure
Ryo Fujii, Future Corporation
Atsushi Fujita, National Institute of Information and Communications Technology
Marco Gaido, Fondazione Bruno Kessler, University of Trento
Baban Gain, Indian Institute of Technology, Patna
Javier Garcia Gilabert, Barcelona Super Computing Center
Sofía García, imaxin|software
Mercedes García-Martínez, Pangeanic
Harritxu Gete, Vicomtech
Jesús González-Rubio, WebInterpret
Isao Goto, Ehime University
Thamme Gowda, Microsoft
Varun Gumma, Microsoft
Kamil Guttmann, Laniqo, Adam Mickiewicz University in Poznań
Jeremy Gwinnup, Air Force Research Laboratory
Anne Göhring, University of Zurich
Nizar Habash, New York University Abu Dhabi

Sami Haq, Dublin City University
Ali Hatami, University of Galway
Xiaoyu He, University of Edinburgh
Jindřich Helcl, Charles University in Prague
John Henderson, Mechanical Learning
Amr Hendy, Microsoft
Shohei Higashiyama, National Institute of Information and Communications Technology
Tosho Hirasawa, OMRON SINIC X Corporation
Miroslav Hrabal, Charles University
Shujian Huang, National Key Laboratory for Novel Software Technology, Nanjing University
Rudali Huidrom, ADAPT Research Centre, Dublin City University
Ifeoluwatayo Ige, Rochester Institute of Technology
Kenji Imamura, National Institute of Information and Communications Technology
Mert Inan, Northeastern University
Jillaphat Jaroenkantasima, Kasetsart University
Advait Joglekar, Indian Institute of Technology Madras
Josef Jon, Charles University
Nisheeth Joshi, Banasthali Vidyapith
Takatomo Kano, NTT
Marzena Karpinska, University of Massachusetts Amherst
Kishore Kashyap, Department of Information Technology , Gauhati University
Kazutaka Kinugawa, NHK Science and Technology Research Laboratories
Mateusz Klimaszewski, Warsaw University of Technology
Sai Koneru, Karlsruhe Institute of Technology
Maarit Koponen, University of Eastern Finland
Elizaveta Korotkova, University of Tartu
Mikołaj Koszowski, ML Research at Allegro
Artur Kot, Allegro.com
Lea Krause, Vrije Universiteit Amsterdam
Parameswari Krishnamurthy, Assistant Professor, IIIT Hyderabad
Mateusz Krubinski, Snowflake
Roland Kuhn, National Research Council of Canada
Ashish Kulkarni, Krutrim
Anoop Kunchukuttan, Microsoft AI and Research
Ali Kuzhuget, tyvan.ru
Wen Lai, Technical University of Munich
Lenin Laitonjam, NIT Mizoram
Ekaterina Lapshinova-Koltunski, University of Hildesheim
Samuel Larkin, National Research Council Canada
Sahinur Rahman Laskar, UPES Dehradun
Alon Lavie, Unbabel/Carnegie Mellon University
Annie Lee, University of Toronto
Gregor Leusch, eBay
William Lewis, University of Washington
Ben Li, Nomura Research Institute, Ltd.
Zhenhao Li, Imperial College London
Baohao Liao, University of Amsterdam
Xixian Liao, Barcelona Supercomputing Center
Lisa Liu, University of California, San Diego
Wuying Liu, Ludong University
Danni Liu, Karlsruhe Institute of Technology

Siyou Liu, University of Macau
Chi-Kiu Lo, National Research Council of Canada
Henrique Lopes Cardoso, University of Porto
Jiaming Luo, Google
Nam Luu, Charles University
Chenyang Lyu, MBZUAI
Andreas Maletti, Universität Leipzig
Bhavitvya Malik, University of Edinburgh
Shushen Manakhimova, German Research Center for Artificial Intelligence (DFKI)
Euan Mcgill, Universitat Pompeu Fabra
Maite Melero, BSC
Yan Meng, University of Amsterdam
Antonio Valerio Miceli Barone, The University of Edinburgh
Nikolay Mikhaylovskiy, NTR Labs / Higher IT School of Tomsk State University
Hideya Mino, NHK Science and Technology Research Laboratories
Sthembiso Mkhwanazi, CSIR
Wafaa Mohammed, University of Amsterdam
Amit Moryossef, Bar-Ilan university, University of Zurich
Yasmin Moslem, ADAPT Centre, Trinity College Dublin
Ananya Mukherjee, International Institute of Information Technology Hyderabad
Aniruddha Mukherjee, Kalinga Institute of Industrial Technology
Kenton Murray, Johns Hopkins University
Jonathan Mutal, Unige
Mathias Müller, University of Zurich
Masaaki Nagata, NTT Inc.
Toshiaki Nakazawa, The University of Tokyo
Subhajit Naskar, Google
Angel Navarro, PRHLT
Prashanth Nayak, KantanAI
Mariana Neves, German Federal Institute for Risk Assessment
Jan Niehues, Karlsruhe Institut of Technology
Takashi Ninomiya, Ehime University
Lydia Nishimwe, Inria
Artur Nowakowski, Laniqo / Adam Mickiewicz University
Perez Ogayo, Carnegie Mellon University
Tsuyoshi Okita, Kyushu institute of technology
Antoni Oliver, Universitat Oberta de Catalunya
Constantin Orasan, University of Surrey
Lucía Ormaechea, Université de Genève
Daniel Ortiz-Martinez, University of Barcelona
Partha Pakray, National Institute of Technology Silchar
Santanu Pal, Wipro
Jianhui Pang, University of Macau
Chanjun Park, Korea University
Pavel Pecina, Charles University
Stephan Peitz, Apple
Sergio Penkale, Lingo24
Frithjof Petrick, AppTek
Pavel Petrushkov, eBay
Sanjita Phijam, National Institute of Technology Silchar
Mārcis Pinnis, Tilde

Mikołaj Pokrywka, Laniqo, Allegro, Adam Mickiewicz University
Jose Pombal, Unbabel
Martin Popel, Charles University, Faculty of Mathematics and Physics, UFAL
Maja Popović, IU University
Matt Post, Microsoft
Pawan Rajpoot, Self
Surangika Ranathunga, Massey University
Vikas Raunak, Microsoft
Pretam Ray, Indian Institute of Technology Kharagpur
Stephen Richardson, Brigham Young University
Matiss Rikters, AIST
Nathaniel Robinson, Johns Hopkins University
Raphael Rubino, UNIGE
Pavel Rychly, NLP Centre, Faculty of Informatics, Masaryk University
Benoît Sagot, Inria
Pramit Sahoo, Indian Institute of Technology Hyderabad
Yusuke Sakai, Nara Institute of Science and Technology
Elizabeth Salesky, Johns Hopkins University
Loitongbam Sanayai Meetei, National Institute of Technology Silchar
Marcelo Sandoval-Castaneda, Toyota Technological Institute at Chicago
Aleix Sant, Barcelona Supercomputing Center
Ashish Sardana, NVIDIA
Hamees Sayed, Indian Institute of Technology Madras
Yves Scherrer, University of Oslo
Florian Schottmann, TextShuttle, ETH Zurich
Rico Sennrich, University of Zurich
Hendra Setiawan, Apple Inc.
Apurva Shah, Google Inc
Lia Shahnazaryan, Paderborn University
Mohammed Shaikheldin, Independent Researcher
Sherrie Shen, University of Edinburgh
Ketaki Shetye, International Institute of Information Technology
Manish Shrivastava, International Institute of Information Technology Hyderabad
Ammon Shurtz, Brigham Young University
Edoardo Signoroni, NLP Centre, Faculty of Informatics, Masaryk University
Thoudam Doren Singh, National Institute of Technology Meghalaya
Huacheng Song, The Hong Kong Polytechnic University
Rui Sousa-Silva, University of Porto - Faculty of Arts and Humanities
Felix Stahlberg, Google Research
Steinthor Steingrimsson, The Arni Magnusson Institute for Icelandic Studies
Steinthór Steingrímsson, The Árni Magnússon Institute for Icelandic Studies
Katsuhito Sudoh, Nara Women's University
Haoxiang Sun, ShanghaiJiaotong University
Eduardo Sánchez, Meta/UCL
George Tambouratzis, ILSP/Athena R.C.
Aleš Tamchyna, Memsource
Shaomu Tan, University of Amsterdam
Gongbo Tang, Beijing Language and Culture University
Jörg Tiedemann, University of Helsinki
Evgeniia Tokarchuk, University of Amsterdam
Isidora Tourni, Boston University

Marco Turchi, Zoom Video Communications
Masao Utiyama, NICT
Takehito Utsuro, University of Tsukuba
Jannis Vamvas, Department of Computational Linguistics, University of Zurich
Dusan Varis, Charles University, Institute of Formal and Applied Linguistics
Raul Vazquez, University of Helsinki
Menan Velayuthan, University of Jaffna
Nitin Venkateswaran, University of Florida
Federica Vezzani, University of Padua
David Vilar, Google
Ekaterina Vylomova, University of Melbourne
Xiaotian Wang, University of Tsukuba
Wei Wang, Apple AI/ML
Taro Watanabe, Nara Institute of Science and Technology
Rachel Wicks, Johns Hopkins University
Guillaume Wisniewski, Universite Paris Cite and LLF
Derek F. Wong, University of Macau
Yulong Wu, The University of Manchester
Minghao Wu, Monash University
Di Wu, University of Amsterdam
Tong Xiao, Northeastern University
Hongzhi Xu, Shanghai International Studies University
Saumitra Yadav, International Institute of Information Technology, Hyderabad
Hao Yang, Huawei Co. Ltd
Kazuki Yano, Tohoku University
Zekai Ye, Harbin Institute of Technology
Hyeongu Yun, LG AI Research
François Yvon, ISIR CNRS and Sorbonne Université
Armel Randy Zebaze Dongmo, Inria
Hui Zeng, Jiaguyi Language Technology Co., Ltd
Chrysoula Zerva, Instituto de Instituto de Telecomunicações, Instituto Superior Técnico, University of Lisbon
Runzhe Zhan, University of Macau
Xuan Zhang, Johns Hopkins University
Dakun Zhang, CHAPSVISION
Min Zhang, Huawei
Shaolin Zhu, Tianjin university
Lichao Zhu, Paris CitÃ© University
Maria Zimina, Paris Diderot – Sorbonne Paris Cité, CLILLAC-ARP (EA 3967)
Maria Zimina-Poirot, ALTAE, UniversitÃ© Paris CitÃ©
Vilém Zouhar, ETH Zurich, Charles University

Table of Contents

Program

Saturday, November 8, 2025

09:00 - 09:10 *Opening Remarks - 20 Years of WMT*

09:10 - 10:30 *Session 1 - Shared Task Overview Papers I*

Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets
Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya and Vilém Zouhar

Findings of the WMT25 Multilingual Instruction Shared Task: Persistent Hurdles in Reasoning, Generation, and Evaluation
Tom Kocmi, Sweta Agrawal, Ekaterina Artemova, Eleftherios Avramidis, Eleftheria Briakou, Pinzhen Chen, Marzieh Fadaee, Markus Freitag, Roman Grundkiewicz, Yupeng Hou, Philipp Koehn, Julia Kreutzer, Saab Mansour, Stefano Perrella, Lorenzo Proietti, Parker Riley, Eduardo Sánchez, Patricia Schmidtova, Mariya Shmatova and Vilém Zouhar

Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help
Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag and Daniel Deutsch

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Session 2 - Shared Task Posters*

11:00 - 12:00 *General Translation Task*

Midheind at WMT25 General Machine Translation Task
Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Kári Steinn Adhalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Thórdharson and Thorvaldur Páll Helgason

A Preliminary Study of AI Agent Model in Machine Translation
Ahrii Kim

Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding
Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo and Kaifu Zhang

Evaluation of QWEN-3 for English to Ukrainian Translation
Cristian Grozea and Oleg Verbitsky

SYSTRAN @ WMT 2025 General Translation Task
Dakun Zhang, Yara Khater, Ramzi Rahli, Anna Rebollo and Josep Crego

Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task
Mao Zheng, Zheng Li, Yang Du, Bingxin Qu and Mingyang Song

From SALAMANDRA to SALAMANDRATA: BSC Submission for WMT25 General Machine Translation Shared Task
Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt, Carlos Escolano and Maite Melero

Instruction-Tuned English to Bhojpuri Neural Machine Translation Using Contrastive Preference Optimization
Kshetrimayum Boynao Singh, Deepak Kumar and Asif Ekbal

SH at WMT25 General Machine Translation Task
Hayate Shiroma

Simple Test Time Scaling for Machine Translation: Kaze-MT at the WMT25 General Translation Task
Shaomu Tan

NTTSU at WMT2025 General Translation Task
Zhang Yin, Hiroyuki Deguchi, Haruto Azami, Guanyu Ouyang, Kosei Buma, Yingyi Fu, Katsuki Chousa and Takehito Utsuro

A Decoding for Machine Translation in LLMs - SRPOL Participation in WMT2025*
Adam Dobrowolski, Paweł Przewłocki, Paweł Przybysz, Marcin Szymański and Dawid Siwicki

In2X at WMT25 Translation Task
Lei Pang, Hanyi Mao, Quanjia Xiao, Chen Ruihan, Jingjun Zhang, Haixiao Liu and Xiangyi Li

CUNI at WMT25 General Translation Task
Josef Jon, Miroslav Hrabal, Martin Popel and Ondřej Bojar

UvA-MT's Participation in the WMT25 General Translation Shared Task
Di Wu, Yan Meng, Maya Konstantinovna Nachesa, Seth Aycock and Christof Monz

AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama Models for Low Resource Machine Translation
Atli Jasonarson and Steinthor Steingrimsson

KIKIS at WMT 2025 General Translation Task
Koichi Iwakawa, Keito Kudo, Subaru Kimura, Takumi Ito and Jun Suzuki

Google Translate's Research Submission to WMT2025
Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Markus Freitag and David Vilar

DLUT and GTCOM's Large Language Model Based Translation System for WMT25
Hao Zong, Chao Bei, Wentao Chen, Conghu Yuan, Huan Liu and Degen Huang

Yandex Submission to the WMT25 General Machine Translation Task
Nikolay Karpachev, Ekaterina Enikeeva, Dmitry Popov, Arsenii Bulgakov, Daniil Panteleev, Dmitrii Ulianov, Artem Kryukov and Artem Mekhraliev

IRB-MT at WMT25 Translation Task: A Simple Agentic System Using an Off-the-Shelf LLM
Ivan Grubišić and Damir Korencic

Improving Low-Resource Japanese Translation with Fine-Tuning and Backtranslation for the WMT 25 General Translation Task
Felipe Fujita and Hideyuki Takada

Multi-agentMT: Deploying AI Agent in the WMT25 Shared Task
Ahrii Kim

Laniqo at WMT25 General Translation Task: Self-Improved and Retrieval-Augmented Translation
Kamil Guttmann, Zofia Rostek, Adrian Charkiewicz, Antoni Solarski, Mikołaj Pokrywka and Artur Nowakowski

Command-A-Translate: Raising the Bar of Machine Translation with Difficulty Filtering
Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent and Ivan Zhang

11:00 - 12:00 *Test Suites Shared Task*

GENDER1PERSON: Test Suite for Estimating Gender Bias of First-person Singular Forms
Maja Popović and Ekaterina Lapshinova-Koltunski

Evaluation of LLM for English to Hindi Legal Domain Machine Translation Systems
Kshetrimayum Boynao Singh, Deepak Kumar and Asif Ekbal

RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation
Rachel Bawden and Benoît Sagot

Automated Evaluation for Terminology Translation Related to the EEA Agreement
Selma Dis Hauksdottir and Steinthor Steingrimsson

Up to Par? MT Systems Take a Shot at Sports Terminology
Einar Sigurdsson, Magnús Magnússon, Atli Jasonarson and Steinthor Steingrimsson

Fine-Grained Evaluation of English-Russian MT in 2025: Linguistic Challenges Mirroring Human Translator Training
Shushen Manakhimova, Maria Kunilovskaya, Ekaterina Lapshinova-Koltunski and Eleftherios Avramidis

11:00 - 12:00 *Translation Quality Evaluation Shared Task*

Tagged Span Annotation for Detecting Translation Errors in Reasoning LLMs
Taemin Yeom, Yonghyun Ryu, Yoonjung Choi and Jinyeong Bak

COMET-poly: Machine Translation Metric Grounded in Other Candidates
Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues and Mrinmaya Sachan

Long-context Reference-based MT Quality Estimation
Sami Haq, Chinonso Osuji, Sheila Castilho, Brian Davis and Thiago Castro Ferreira

Evaluating WMT 2025 Metrics Shared Task Submissions on the SSA-MTE African Challenge Set
Senyu Li, Felermino Dario Mario Ali, Jiayi Wang, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, Colin Cherry and David Ifeoluwa Adelani

Nvidia-Nemo's WMT 2025 Metrics Shared Task Submission
Brian Yan, Shuoyang Ding, Kuang-Da Wang, Siqi Ouyang, Oleksii Hrinchuk, Vitaly Lavrukhin and Boris Ginsburg

GEMBA V2: Ten Judgments Are Better Than One
Marcin Junczys-Dowmunt

CUNI and Phrase at WMT25 MT Evaluation Task
Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie and Craig Stewart

MSLC25: Metric Performance on Low-Quality Machine Translation, Empty Strings, and Language Variants
Rebecca Knowles, Samuel Larkin and Chi-Kiu Lo

MetricX-25 and GemSpanEval: Google Translate Submissions to the WMT25 Evaluation Shared Task
Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang and Markus Freitag

HW-TSC's Submissions to the WMT 2025 Segment-level Quality Score Prediction Task
Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Xiaoyu Chen and Hao Yang

UvA-MT at WMT25 Evaluation Task: LLM Uncertainty as a Proxy for Translation Quality
Di Wu and Christof Monz

Submission for WMT25 Task 3
Govardhan Padmanabhan

RankedCOMET: Elevating a 2022 Baseline to a Top-5 Finish in the WMT 2025 QE Task
Sujal Maharjan and Astha Shrestha

Quality-Informed Segment-Level Error Correction Using Natural Language Explanations from xTower and Large Language Models
Prashant Sharma

TASER: Translation Assessment via Systematic Evaluation and Reasoning
Monishwaran Maheswaran, Marco Carini, Christian Federmann and Tony Diaz

12:00 - 14:00 *Lunch Break*

Saturday, November 8, 2025 (continued)

14:00 - 15:00 *Session 3 - Invited Talk by Longyue Wang (Alibaba)*

15:00 - 16:00 *Coffee Break*

16:00 - 17:30 *Session 4 - Research Papers Talk Session*

An Empirical Analysis of Machine Translation for Expanding Multilingual Benchmarks
Sara Rajaee, Rochelle Choenni, Ekaterina Shutova and Christof Monz

Cross-lingual Human-Preference Alignment for Neural Machine Translation with Direct Quality Optimization
Kaden Uhlig, Joern Wuebker, Raphael Reinauer and John Denero

Audio-Based Crowd-Sourced Evaluation of Machine Translation Quality
Sami Haq, Sheila Castilho and Yvette Graham

Meaningful Pose-Based Sign Language Evaluation
Zifan Jiang, Colin Leong, Amit Moryossef, Oliver Cory, Maksym Ivashechkin, Neha Tarigopula, Biao Zhang, Anne Göhring, Annette Rios, Rico Sennrich and Sarah Ebling

Context Is Ubiquitous, but Rarely Changes Judgments: Revisiting Document-Level MT Evaluation
Ahrii Kim

GIIFT: Graph-guided Inductive Image-free Multimodal Machine Translation
Jiafeng Xiong and Yuting Zhao

Sunday, November 9, 2025

09:10 - 10:30 *Session 5 - Shared Task Overview Papers II*

Findings of the WMT 2025 Shared Task on Model Compression: Early Insights on Compressing LLMs for Machine Translation
Marco Gaido, Roman Grundkiewicz, Thamme Gowda and Matteo Negri

Findings of the WMT 2025 Shared Task of the Open Language Data Initiative
David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell and Philipp Koehn

Findings of the WMT 2025 Shared Task LLMs with Limited Resources for Slavic Languages: MT and QA
Shu Okabe, Daryna Dementieva, Marion Di Marco, Lukas Edman, Katharina Haemmerl, Marko Měškank, Anita Hendrichowa and Alexander Fraser

Findings of the First Shared Task for Creole Language Machine Translation at WMT25
Nathaniel Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Kenton Murray, Raj Dabre, Andre Coy and Heather Lent

Findings of WMT 2025 Shared Task on Low-resource Indic Languages Translation
Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das and Riyanka Manna

Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs
Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay and Pinzhen Chen

10:30 - 11:00 *Coffee Break*

09:10 - 10:30 *Session 5 - Shared Task Posters II*

11:00 - 12:00 *Model Compression Shared Task*

Vicomtech@WMT 2025: Evolutionary Model Compression for Machine Translation
David Ponce, Harritxu Gete and Thierry Etchegoyhen

Iterative Layer Pruning for Efficient Translation Inference
Yasmin Moslem, Muhammad Hazim Al Farouq and John Kelleher

Sunday, November 9, 2025 (continued)

11:00 - 12:00 *Open Language Data Initiative*

Expanding the WMT24++ Benchmark with Rumantsch Grischun, Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader
Jannis Vamvas, Ignacio Pérez Prat, Not Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna Rutkiewicz and Rico Sennrich

A French Version of the OLDI Seed Corpus
Malik Marmonier, Benoît Sagot and Rachel Bawden

Bringing Ladin to FLORES+
Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto and Paolo Anvidalfarei

Correcting the Tamazight Portions of FLORES+ and OLDI Seed Datasets
Alp Oktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja and Farida Boudichat

Filling the Gap for Uzbek: Creating Translation Resources for Southern Uzbek
Mukhammadsaid Mamasaidov, Azizullah Aral, Abror Shopulatov and Mironshoh Inomjonov

The Kyrgyz Seed Dataset Submission to the WMT25 Open Language Data Initiative Shared Task
Murat Jumashev, Alina Tillabaeva, Aida Kasieva, Turgunbek Omurkanov, Akylai Musaeva, Meerim Emil Kyzy, Gulaiym Chagataeva and Jonathan Washington

SMOL: Professionally Translated Parallel Data for 115 Under-represented Languages
Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Moussa Doumbouya, Djibrila Diane, Baba Mamadi Diane, Solo Farabado, Edoardo Ferrante, Alessandro Guasoni, Mamadou Keita, Sudhamoy Debbarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, Sina Ahmadi, Mingfei Liu and Jonathan Eng

Improved Norwegian Bokmål Translations for FLORES
Petter Mæhlum, Anders Næss Evensen and Yves Scherrer

11:00 - 12:00 *Slavic Languages Shared Task*

NRC Systems for the WMT2025-LRSL Shared Task
Samuel Larkin, Chi-Kiu Lo and Rebecca Knowles

TartuNLP at WMT25 LLMs with Limited Resources for Slavic Languages Shared Task
Taido Purason and Mark Fishel

JGU Mainz's Submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA
Hossain Shaikh Saadi, Minh Duc Bui, Mario Sanz-Guerrero and Katharina Von Der Wense

11:00 - 12:00 *Creole Language Translation Shared Task*

Krey-All WMT 2025 CreoleMT System Description: Language Agnostic Strategies for Low-Resource Translation
Ananya Ayasi

EdinHelsOW WMT 2025 CreoleMT System Description: Improving Lusophone Creole Translation through Data Augmentation, Model Merging and LLM Post-editing
Jacqueline Rowe, Ona De Gibert, Mateusz Klimaszewski, Coleman Haley, Alexandra Birch and Yves Scherrer

KozKreolMRU WMT 2025 CreoleMT System Description: Koz Kreol: Multi-Stage Training for English–Mauritian Creole MT
Yush Rajcoomar

JHU WMT 2025 CreoleMT System Description: Data for Belizean Kriol and French Guianese Creole MT
Nathaniel Robinson

WMT 2025 CreoleMT Systems Description : Martinican Creole and French
Ludovic Mompelat

11:00 - 12:00 *Low-Resource Indic Language Translation Shared Task*

JU-NLP: Improving Low-Resource Indic Translation System with Efficient LoRA-Based Adaptation
Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das and Sivaji Bandyopadhyay

An Attention-Based Neural Translation System for English to Bodo
Subhash Wary, Birhang Borgoyary, Akher Ahmed, Mohanji Sah and Apurbalal Senapati

Tackling Low-Resource NMT with Instruction-Tuned LLaMA: A Study on Kokborok and Bodo
Deepak Kumar, Kshetrimayum Boynao Singh and Asif Ekbal

DELAB-IIITM WMT25: Enhancing Low-Resource Machine Translation for Manipuri and Assamese
Dingku Oinam and Navanath Saharia

Transformers: Leveraging OpenNMT and Transfer Learning for Low-Resource Indian Language Translation
Bhagyashree Wagh, Harish Bapat, Neha Gupta and Saurabh Salunkhe

RBG-AI: Benefits of Multilingual Language Models for Low-Resource Languages
Barathi Ganesh Hb and Michal Ptaszynski

ANVITA : A Multi-pronged Approach for Enhancing Machine Translation of Extremely Low-Resource Indian Languages
Sivabhavani J, Daneshwari Kankanwadi, Abhinav Mishra and Biswajit Paul

DoDS-IITPKD:Submissions to the WMT25 Low-Resource Indic Language Translation Task
Ontiwell Khongthaw, G.l. Salvin, Shrikant Budde, Abigairl Chigwededza, Dhruvadeep Malkar and Swapnil Hingmire

A Preliminary Exploration of Phrase-Based SMT and Multi-BPE Segmentations through Concatenated Tokenised Corpora for Low-Resource Indian Languages
Saumitra Yadav and Manish Shrivastava

AkibaNLP-TUT: Injecting Language-Specific Word-Level Noise for Low-Resource Language Translation
Shoki Hamada, Tomoyosi Akiba and Hajime Tsukada

BVSLP: Machine Translation Using Linguistic Embellishments for IndicMT Shared Task 2025
Nisheeth Joshi, Palak Arora, Anju Krishnia, Riya Lonchenpa and Mhasilenuo Vizo

TranssionMT's Submission to the Indic MT Shared Task in WMT 2025
Zebiao Zhou, Hui Li, Xiangxun Zhu and Kangzhen Liu

11:00 - 12:00 *Terminology Translation Shared Task*

Laniqo at WMT25 Terminology Translation Task: A Multi-Objective Reranking Strategy for Terminology-Aware Translation via Pareto-Optimal Decoding
Kamil Guttmann, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka and Artur Nowakowski

Fine-tuning NMT Models and LLMs for Specialised EN-ES Translation Using Aligned Corpora, Glossaries, and Synthetic Data: MULTITAN at WMT25 Terminology Shared Task
Lichao Zhu, Maria Zimina-Poirot, Stephane Patin and Cristian Valdez

Contextual Selection of Pseudo-terminology Constraints for Terminology-aware Neural Machine Translation in the IT Domain
Benjamin Pong

IRB-MT at WMT25 Terminology Translation Task: Metric-guided Multi-agent Approach
Ivan Grubišić and Damir Korencic

Terminology-Constrained Translation from Monolingual Data Using GRPO
Javier Garcia Gilabert, Carlos Escolano, Xixian Liao and Maite Melero

It Takes Two: A Dual Stage Approach for Terminology-Aware Translation
Akshat Jaswal

12:00 - 14:00 *Lunch Break*

14:00 - 15:30 *Session 7 - Research Paper Boaster Session*

15:30 - 16:00 *Coffee Break*

16:00 - 17:00 *Session 8 - Research Paper Poster Session*

Specification-Aware Machine Translation and Evaluation for Purpose Alignment
Yoko Kayano and Saku Sugawara

OpenWHO: A Document-Level Parallel Corpus for Health Translation in Low-Resource Languages
Raphael Merx, Hanna Suominen, Trevor Cohn and Ekaterina Vylomova

Factors Affecting Translation Quality in In-context Learning for Multilingual Medical Domain
Jonathan Mutal, Raphael Rubino and Pierrette Bouillon

Context Is Ubiquitous, but Rarely Changes Judgments: Revisiting Document-Level MT Evaluation
Ahrii Kim

Character-Aware English-to-Japanese Translation of Fictional Dialogue Using Speaker Embeddings and Back-Translation
Ayuna Nagato and Takuya Matsuzaki

DTW-Align: Bridging the Modality Gap in End-to-End Speech Translation with Dynamic Time Warping Alignment
Abderrahmane Issam, Yusuf Can Semerci, Jan Scholtes and Gerasimos Spanakis

Targeted Source Text Editing for Machine Translation: Exploiting Quality Estimators and Large Language Models
Hyuga Koretaka, Atsushi Fujita and Tomoyuki Kajiwara

Self-Retrieval from Distant Contexts for Document-Level Machine Translation
Ziqian Peng, Rachel Bawden and François Yvon

Using Encipherment to Isolate Conditions for the Successful Fine-tuning of Massively Multilingual Translation Models
Carter Louchheim, Denis Sotnichenko, Yukina Yamaguchi and Mark Hopkins

Translate, Then Detect: Leveraging Machine Translation for Cross-Lingual Toxicity Classification
Samuel Bell, Eduardo Sánchez, David Dale, Pontus Stenetorp, Mikel Artetxe and Marta R. Costa-Jussà

Feeding Two Birds or Favoring One? Adequacy–Fluency Tradeoffs in Evaluation and Meta-Evaluation of Machine Translation
Behzad Shayegh, Jan-Thorsten Peter, David Vilar, Tobias Domhan, Juraj Juraska, Markus Freitag and Lili Mou

DocHPLT: A Massively Multilingual Document-Level Translation Dataset
Dayyán O'brien, Bhavitvya Malik, Ona De Gibert, Pinzhen Chen, Barry Haddow and Jörg Tiedemann

SONAR-SLT: Multilingual Sign Language Translation via Language-Agnostic Sentence Embedding Supervision
Yasser Hamidullah, Shakib Yazdani, Cennet Oguz, Josef Van Genabith and Cristina España-Bonet

GAMBIT+: A Challenge Set for Evaluating Gender Bias in Machine Translation Quality Estimation Metrics
George Filandrianos, Orfeas Menis Mastromichalakis, Wafaa Mohammed, Giuseppe Attanasio and Chrysoula Zerva

Implementing and Evaluating Multi-source Retrieval-Augmented Translation
Tommi Nieminen, Jörg Tiedemann and Sami Virpioja

A Cross-Lingual Perspective on Neural Machine Translation Difficulty
Esther Ploeger, Johannes Bjerva, Jörg Tiedemann and Robert Oestling

An Empirical Analysis of Machine Translation for Expanding Multilingual Benchmarks

Sara Rajaee[*] [♢] **Rochelle Choenni**[*] [♠] **Ekaterina Shutova**[♠] **Christof Monz**[♢]
[♠]ILLC, University of Amsterdam, the Netherlands
[♢]Language Technology Lab, University of Amsterdam, the Netherlands
{s.rajaee, r.m.v.k.choenni, e.shutov, c.monz}@uva.nl

Abstract

The rapid advancement of large language models (LLMs) has introduced new challenges in their evaluation, particularly for multilingual settings. The limited evaluation data are more pronounced in low-resource languages due to the scarcity of professional annotators, hindering fair progress across languages. In this work, we systematically investigate the viability of using machine translation (MT) as a proxy for evaluation in scenarios where human-annotated test sets are unavailable. Leveraging a state-of-the-art translation model, we translate datasets from four tasks into 198 languages and employ these translations to assess the quality and robustness of MT-based multilingual evaluation under different setups. We analyze task-specific error patterns, identifying when MT-based evaluation is reliable and when it produces misleading results. Our translated benchmark reveals that current language selections in multilingual datasets tend to overestimate LLM performance on low-resource languages. We conclude that although machine translation is not yet a fully reliable method for evaluating multilingual models, overlooking its potential means missing a valuable opportunity to track progress in non-English languages.

1 Introduction

Large-scale evaluation of multilingual language models (MLMs) across hundreds of languages has remained a persistent challenge as the existing benchmarks cover a subset, and mostly high-resource, of languages (Singh et al., 2024, 2025). Although human-translated (HT) evaluation sets offer accurate and reliable evaluation of MLMs, creating them for hundreds of languages is costly, time-consuming, and sometimes impossible. Consequently, tracking the progress of MLMs in most languages has lagged behind. Moreover, multilingual benchmarks often cover different subsets of languages, and such inconsistent evaluation setups create a fragmented picture of MLM capabilities (Liang et al., 2020; Hu et al., 2020; Ruder et al., 2021).

Recent advances in machine translation (MT) have emerged as a promising solution to these challenges of multilingual evaluation. Prior studies have shown that state-of-the-art MT systems achieve human-level translation quality in certain high-resource languages (Kocmi et al., 2023, 2024). While MT has been used to construct evaluation sets (Chen et al., 2024), to our knowledge, there has been no systematic study of the viability of MT-based evaluations across different evaluation setups and diverse downstream tasks in nearly 200 languages. Moreover, although previous work has shown the limitations of MT models, such as translation artifacts and stylistic shifts (Park et al., 2024; Wang et al., 2023), there has been no comprehensive analysis of the potential risks of using MT-based datasets for MLM evaluation.

In this work, and by considering recent MT advances, we investigate the viability of using MT for large-scale multilingual evaluation where human-annotated data is unavailable. Using four popular multilingual tasks, we translate their test sets into 198 languages using NLLB ("No Language Left Behind"), which officially supports 200 languages (NLLB Team et al., 2022). We then compare three MLMs' performance, i.e., XLM-R (both base and large versions) (Conneau et al., 2020), BLOOMz (Muennighoff et al., 2022), and AYA-101 (Üstün et al., 2024b), on these MT-based test sets against their HT equivalents, using multiple evaluation setups (zero-shot fine-tuning and zero-shot prompting) and both accuracy and rank correlation as metrics. Beyond overall performance comparisons, we analyze how translation quality relates to evaluation results, detecting MT-specific error types using an LLM-as-a-judge framework in the selected tasks. Finally, we assess whether

[*]Equal contribution.

Proceedings of the Tenth Conference on Machine Translation, pages 1–30
November 8-9, 2025

current multilingual benchmarks misrepresent the performance of MLMs when limited to small language subsets. More specifically, we try to answer the following research questions:

- To what extent does the use of machine translation affect MLM performance estimates?
- To what extent is MLM performance on MT-based evaluations influenced by translation quality?
- What major translation error types occur in cases where MLMs succeed on human translations but fail on machine translations?
- To what extent do existing multilingual benchmarks misrepresent MLM performance?

Across tasks and models, MT-based evaluations yield results that are highly correlated with HT-based evaluations (average Spearman's 0.95), with only small average accuracy differences (<1.5 points), Table 3 and 4. Furthermore, translation quality shows a moderate positive correlation with performance differences, suggesting that MT reliability depends partly on the underlying MT system's accuracy and the sensitivity of the downstream task to the translation quality, Table 5. Our error analysis reveals that lower performance on MT data is often linked to lexical mistranslations and subtle semantic shifts, though major meaning-altering errors are relatively rare. We show that LLM-as-a-judge might serve as a proxy for filtering low-quality translation examples, thereby improving the reliability of MT-based evaluations. Finally, we find that restricting evaluation to the language subsets in the widely-used benchmarks underestimates the performance of MLMs on high-resource languages and overestimates their performance on low-resource languages by up to 4 accuracy points compared to evaluation across 198 languages.

2 Related work

Multilingual large language models (LLMs) are rapidly expanding their language coverage, reaching to hundreds of languages (Üstün et al., 2024a; Yang et al., 2025). Yet, benchmarks have struggled to keep pace, often covering only a fraction of these languages (e.g., Global MMLU covers 42 languages (Singh et al., 2025), whereas models like Gemma 3 support over 140 (Gemma Team et al., 2025)). This imbalance makes it difficult to measure progress fairly and consistently across languages. On the other hand, creating human-annotated benchmarks for every language is impractical due to high costs and limited expertise. In this regard, machine translation is an attractive potential alternative to expand existing multilingual datasets at scale. Although previous studies have shown MT systems may introduce challenges such as hallucinations and translationese artifacts (Artetxe et al., 2020; Wang and Sennrich, 2020; Zhang and Toral, 2019), recent improvements, especially for mid and low-resource languages, have considerably enhanced translation quality (Ranathunga et al., 2023). Thus, leveraging advanced MT systems is an efficient way to expand the language coverage of benchmarks more quickly. This approach opens the door to more comprehensive and equitable evaluations, better reflecting the multilingual capabilities of LLMs. While Thellmann et al. (2024) also investigate whether machine translated benchmarks can reliably assess model performance across languages, they limit the scope of their study to 20 European languages. Since European languages are already overrepresented in existing NLP resources and benchmarks (Asai et al., 2022), while many other languages remain severely underrepresented (Joshi et al., 2020), we instead scale our evaluation to 198 languages to provide a more balanced assessment.

3 Methods

In this paper, we evaluate multilingual LLMs on four tasks across 198 languages. To this end, we create MT test data using the NLLB model and assess performance under two evaluation paradigms: zero-shot testing and zero-shot prompting. In this section we provide a comprehensive overview of our methods used for large-scale multilingual evaluation. In Section 4, we will explain which methods were used to generate the MT test data itself.

3.1 Tasks and datasets

We massively scale the evaluation of LLMs on four datasets.

XNLI The Cross-Lingual Natural Language Inference (XNLI) dataset (Conneau et al., 2018) contains premise-hypothesis pairs labeled with: 'entailment', 'neutral', or 'contradiction' in 15 languages.

The original pairs come from English and the

test sets were human translated into the other languages.

PAWS-X The Cross-Lingual Paraphrase Adversaries from Word Scrambling (PAWS-X) dataset (Yang et al., 2019) requires the model to determine whether two sentences are paraphrases of one another. The parallel test data has been provided in 7 languages. To create this dataset, a subset of the PAWS development and test sets (Zhang et al., 2019) was professionally translated from English to 6 other languages.

XCOPA The Cross-lingual Choice of Plausible Alternatives (Ponti et al., 2020) evaluates common-sense reasoning in 11 languages. The samples contain a premise and question paired with two answer choices from which the model can select. The dataset is manually translated from English into 11 other languages.

XStorycloze The Cross-lingual Storycloze (Lin et al., 2021) proposes a common-sense reasoning task in 11 languages, in which the model predicts which one of two story endings is the most likely to follow after a given short story. The Storycloze dataset was professionally translated from English into 10 other languages.

3.2 Multilingual language models

For the large-scale evaluation, we focus on the base and large version of XLM-R (Conneau et al., 2020) pre-trained on 100 languages as it is one of the most popular MLMs. In addition, we report scores from BLOOMz 7.1b (Scao et al., 2022) and AYA-101 13b (AYA)(Üstün et al., 2024b). BLOOMz is trained on 46 languages and AYA on 101 languages. Moreover, both models are further instruction-tuned on a mixture of prompts in different languages. Given that the PAWS-X dataset was included during instruction-tuning, we evaluate BLOOMz and AYA on the held out datasets only, i.e., XNLI, XCOPA and XStorycloze. Note that BLOOMz and AYA were selected over other MLMs, such as Llama (Touvron et al., 2023), as they are the largest publicly available and explicitly MLMs (i.e., statistics on the pretraining data distribution across languages is publicly available).

3.3 Selection of test languages

For our selection of test languages, there are two constraining factors: (1) the language has to be covered by the NLLB-200 translation model and FLORES-200 dataset, and (2) the script of the language needs to have been seen during the pretraining of the model. We then separately filter out the test languages by unseen scripts for each model. This leaves us with 196, 168, and 198 test languages for XLM-R, BLOOMz, and AYA, respectively, see Appendix J for the complete lists. Note that the number of compatible languages is lower for BLOOMz as it has seen fewer writing scripts during pretraining.

	Unseen	Low	Mid	High	Total
% of data	0	>0 and <0.1	$\geq$ 0.1 and <1	$\geq$1	
XLM-R	106	30	34	26	196
BLOOMz	131	21	7	9	168
AYA	103	57	25	13	198

Table 1: Number of languages categorized as high, mid, low, and unseen languages when looking at the percentage of seen pretraining data of the respective LMs.

Resource categorization by pretraining distribution Moreover, we categorize the test languages for each model separately based on the percentage of total data that they contributed during pretraining. In Table 1, we report the percentage thresholds used for our categorization and the resulting number of test languages for each category and model. As the pretraining data coverage is reported in numbers of GB for XLM-R, we convert these scores to percentages of the full pretraining data. For BLOOMz, we use the reported language distribution numbers [1], and for AYA, we consider mT5 pretraining data distribution as it is utilized as the base model for AYA.

3.4 Evaluation settings

Zero-shot testing We fine-tune XLM-R on the entire training set for each task in English. We then use our fine-tuned model for zero-shot testing in the test languages. We fine-tune our model using the HuggingFace Library, see Appendix A for details.

Zero-shot prompting BLOOMz and AYA are instruction-tuned on multiple classification tasks, thus we test these models out-of-the-box in a zero-shot prompting set up. This has the benefit that dataset artifacts, which are commonly known to be leveraged during fine-tuning, cannot be learned (McCoy et al., 2020). As BLOOMz and AYA fail to predict a third option for XNLI (neutral), we report results on a binarized version of the

[1] https://huggingface.co/bigscience/bloom

metric	High	Mid	Low	Unseen	Ave.	Med.
NLLB-Distil						
chrF++	49.34	46.92	43.22	37.12	44.15	45.07
NLLB-3.3B						
chrF++	52.5	50.8	46.1	40.3	44.5	45.5

Table 2: The quality of the MT system across high, mid, and low-resource languages using chrF++ based on the XLM-R model's categorization.

task by aggregating sentences with the 'neutral' and 'contradiction' labels into one class and making the model predict entailment or not. Moreover, the instructions are given in English. see Appendix A for the prompts per task.
Finally, note that our goal is not to compare performance of BLOOMz and AYA to XLM-R but rather to test the reliability of machine-translated test sets in two popular evaluation paradigms.

Evaluation metric As the automatic evaluation metric for testing our MT quality, we use the chrF++ (Popović, 2017). This metric calculates the character and word n-gram overlap between the machine and human reference translations. It is a tokenization-independent metric aligning better with human judgments for morphologically-rich languages compared to BLEU (Tan et al., 2015; Kocmi et al., 2021; Briakou et al., 2023). We also employ the LLMs-as-a-judge technique to evaluate translation quality and analyze the types of errors that occur in translation. For the evaluation, we use the FLORES-200 dataset (NLLB Team et al., 2022), which includes human-translated data for 200 languages, and select 100 sentences from each language.

4 Machine translating test data

For machine translation, we employ the NLLB model covering 202 languages (NLLB Team et al., 2022). Through extensive analysis, Zhu et al. (2024) have shown that NLLB performance surpasses other powerful LLMs such as ChatGPT (OpenAI., 2022) and GPT-4 (Achiam et al., 2023) when translating out of English ($En \Rightarrow$ `tgt` setting). NLLB also demonstrates a minimal difference to the closed source system, Google translate[2] on the Flores-101 dataset. To gain a better understanding of the role of the MT system on the translated data quality and, consequently, its performance on downstream tasks, we experiment with two NLLB versions. We choose the distill NLLB with 600M parameters and the 3.3B NLLB model using greedy sampling. For each example, we translate each sentence (e.g., SENTENCE1 and SENTENCE2 in PAWS-X) separately. In Appendix B, we provide some of the lessons learned from our MT experiments.

In Table 2, we report the average chrF++ scores for translations obtained with the NLLB-Distil and NLLB-3.3B models on the dev set of FLORES-200 dataset including human translation data in 200 languages (NLLB Team et al., 2022). We categorize languages by XLM-R resource ranking (high, mid, low, and unseen) and observe that NLLB-3.3B consistently performs better than the 600M version across all categories. Thus, while the smaller model has been added for analysis purposes, we use the NLLB-3.3B for translation throughout the paper unless stated otherwise. Moreover, we confirm that our scores for all languages are among the best performance of SOTA multilingual MT systems (Bapna et al., 2022; NLLB Team et al., 2022).

5 Evaluating the reliability of MT data

In this section, we study the opportunities and challenges of using MT to extend multilingual benchmarks and assess model performance across a broader range of languages from two perspectives, i.e. the performance gap when measuring on human versus machine translated data and the quality of the MT data itself. Thus, in Section 5.1, we first study the degree to which machine translation alters the reliability and validity of performance assessments for LLMs. Then, in Section 5.2, we evaluate the quality of the machine translated data itself through automatic metrics and LLM-based judgments, both to analyze error patterns and to serve as a proxy for filtering low-quality translated data for evaluation.

5.1 Evaluation gaps between MT and HT

5.1.1 Average performance changes

The key motivation for using machine translation in expanding multilingual benchmarks is to reduce costly and time-consuming human annotation. However, this raises an important question: *Does relying on machine-translated data compromise the validity of LLMs evaluations?* If model performance on MT data is substantially different from that on human translated data, evaluation outcomes

[2] https://translate.google.com/

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XLM-R																									
XCOPA	-	-	-	-	-	69/73	-	-	-	50/57	76/69	75/73	-	-	-	49/59	-	66/58	64/68	-	69/66	71/66	-	68/70	72/73
XStoryCloze	80/80	-	-	-	84/84	-	79/79	-	78/79	-	88/87	-	-	-	71/70	-	84/85	75/76	-	76/75	-	-	-	-	87/86
XNLI	78/78	82/82	82/82	81/80	83/78	-	-	82/83	75/80	-	-	-	-	-	-	-	79/82	70/74	-	-	76/76	77/79	71/78	78/81	79/74
PAWS-X	-	-	90/91	-	91/91	-	-	91/90	-	-	-	-	81/84	81/83	-	-	-	-	-	-	-	-	-	83/82	
BLOOMz																									
XCOPA	-	-	-	-	-	52/52	-	-	-	N/A	78/78	62/65	-	-	-	51/52	-	60/63	75/72	-	N/A	50/50	-	80/77	71/67
XStoryCloze	88/88	-	-	-	91/91	-	84/78	-	85/84	-	91/90	-	-	-	54/52	-	73/73	79/79	-	74/73	-	-	-	-	70/70
B-NLI	71/72	66/68	69/68	65/66	73/74	-	-	72/73	70/72	-	-	-	-	-	-	-	69/70	70/71	-	-	N/A	68/70	68/70	72/73	74/72
AYA																									
XCOPA	-	-	-	-	-	87/84	-	-	-	82/83	87/87	88/88	-	-	-	56/56	-	79/83	86/83	-	84/82	86/85	-	85/84	86/84
XStoryCloze	95/92	-	-	-	94/94	-	83/75	-	93/91	-	91/87	-	-	-	94/86	-	90/82	93/89	-	93/88	-	-	-	-	95/90
B-NLI	78/79	79/79	78/78	78/78	79/80	-	-	79/80	75/75	-	-	-	-	-	-	-	79/79	74/75	-	-	79/79	79/79	74/75	76/77	77/77

Table 3: The (%) accuracy of the models on the human translated (original)/our machine translated datasets. Darker colors indicate bigger gaps between the models' performance on human and machine-translated data.

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
Pearson corr.	95.3	69.3	93.3	98.0	85.0	97.8	98.7	98.0	95.3	90.7
Spearman rank corr.	77.5	82.4	82.6	98.8	89.0	95.1	97.3	81.8	92.5	77.0

Table 4: Pearson and Spearman rank correlation between the performance on MLMs on human-translated (HT)(original data) and machine-translated (MT) data (ours). The high correlations indicate that the performance of MLMs on both machine and human-translated data is similar.

may become unreliable or biased.

To answer this question, we compare model performance on machine and human translated data. We evaluate all models using the same setups and report accuracy scores on machine and human translated data in Table 3.[3] Our results show that the difference in performance on machine and human-translated data is small (on average, about 2.6%). Moreover, this trend is consistent across all models. In some instances, the models achieve slightly higher performance on machine-translated data. We hypothesize that this may be due to the translationese effect. Conversely, in other cases, we observe a drop in performance, which might be caused by low-quality translations. We investigate this issue in Section 5.2, where we explore methods such as LLM-based filtering to detect and reduce the impact of poor quality MT outputs.

5.1.2 Ranking consistency

Building on the finding that accuracy scores between machine-translated and human-translated data are closely aligned, we further assess the reliability of MT-based evaluation by examining the consistency of model rankings, i.e. whether the relative performance of a model across different languages remains stable. Specifically, we want to see if the accuracy differences observed earlier affect which languages a model performs better or worse on. To measure this, we compute the average Pearson and Spearman rank correlations between model performances on human and machine-translated datasets (see Table 4).

The high correlation across all tasks shows that the rank order of the models' performance across different languages using human and machine-translated data is almost the same. This demonstrates that machine translated data could reliably be used to assess the relative differences in model performance across languages.

5.2 The impact of MT quality

Moreover, we analyze whether the quality of machine translation, as measured by chrF++, correlates with LLM performance. In other words, what matters in this experiment is the consistency of the correlations between human- vs. machine-translated data. Table 5 summarizes the numerical results. The results show no significant correlation in either case. That is, higher MT quality (as indicated by better chrF++ scores) does not systematically correspond to higher model performance. The pattern suggests that small variations in MT quality, as captured by the automatic metric, do not strongly influence LLM accuracy. This further supports the robustness of using MT for multilingual evaluation, although it also highlights that chrF++ may not fully capture the aspects of translation quality that affect downstream performance.

[3] Results of XLM-R base are reported in the appendix.

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
Pearson Corr.										
chrF++ vs. HT	27.4	10.5	89.3	3.3	-10.4	-10.4	64.5	41.0	8.2	-18.5
chrF++ vs. MT	30.3	66.2	96.5	8.8	14.5	2.9	66.7	52.1	16.3	11.3
Spearman rank Corr.										
chrF++ vs. HT	29.6	24.4	65.7	16.4	5.7	5.7	68.1	22.7	8.1	-34.6
chrF++ vs. MT	22.6	48.1	82.7	22.4	18.3	14.7	75.8	55.5	26.1	22.4

Table 5: Average Pearson and Spearman rank correlation between chrf++ scores and human translated (HT) (original data) and machine-translated data (MT) (ours).

Source (en)	**Translation (es)**	**Error type (span)**	**Severity**
So I'm not really sure why. I am certain as to the reason why.	Así que no estoy muy seguro de por qué. Estoy seguro de la razón.	Accuracy/omission ("as to") Fluency/redundancy ("de por qué")	Major Minor
I'm covering the same stuff. I'm talking about the same things they did.	Estoy cubriendo las mismas cosas. Estoy hablando de las mismas cosas que ellos hicieron.	Accuracy/mistranslation ("que ellos hicieron") Fluency/grammar ("Estoy cubriendo")	Major Minor

Table 6: Examples of the MQM framework output for detected errors in translated XNLI test examples.

5.2.1 LLM-as-a-Judge Evaluation

In the earlier experiments, we observed that for certain tasks and languages, model performance on MT data was lower than on HT data. We hypothesize that these drops are often linked to lower translation quality. Automatic metrics such as chrF++, while useful, may be overly sensitive to minor deviations that are not critical for the downstream task, resulting in misleadingly low scores.

To better assess translation quality, we employ an LLM-as-a-judge approach. Specifically, we use it to (1) identify and categorize translation errors in examples where the model makes correct predictions on human translations but fails on machine-translated versions, and (2) evaluate its potential as a filtering mechanism to remove low-quality translations from translated evaluation sets.

To this end, we adopt the Multidimensional Quality Metrics (MQM) framework from Freitag et al. (2021). We prompt the LLM evaluator to follow the evaluation guidelines for the translation task to detect and classify translation errors into major and minor categories. The LLM-based evaluation provides the error span, error type, and error classification. Major errors (true errors) are generally easier to detect, whereas minor errors often stem from minor imperfections in translations, see Table 6 for an example.[4] As our LLM-as-a-judge, we use the Gemma 3 27B model (Gemma Team et al., 2025), covering over 140 languages, in an in-context learning setup to evaluate translation quality.[5]

In Table 7, we report the average number of major errors for all predictions (Ave. #Err.) and for the subset where predictions switch from correct on HT to wrong on MT (C→W). While the average error rates for these two categories are close, C→W shows slightly higher values across most languages, suggesting that these degraded translations can be systematically flagged. Moreover, consistently across all tasks and languages, the most frequent translation error type is *accuracy*, including a mistranslation error or omission from the source sentence. Using this observation, we filter out examples with more than 2 major errors, and evaluate the performance of the AYA model on the higher-quality examples. In Table 8, we present the percentage of gap reduction between the models' performance on HT and MT. Based on the results, in most cases, after removing low-quality MT data, the performance gap between MT and HT data is reduced up to 100%. Our results indicate that integrating LLM-based quality checks into MT-based evaluation pipelines can help reduce evaluation noise and improve the reliability of us-

[4] See Table 2 in Freitag et al. (2021) for an overview of the error categorizations.

[5] Please refer to the appendix for the experimental setups.

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XCOPA																									
Ave. #Err.	-	-	-	-	-	0.98	-	-	-	1.50	0.82	0.67	-	-	-	2.30	-	1.19	1.29	-	1.21	0.81	-	0.87	1.13
C→W	-	-	-	-	-	1.10	-	-	-	1.57	1.04	0.57	-	-	-	2.52	-	1.39	1.11	-	1.41	0.90	-	0.91	1.37
XStoryCloze																									
Ave. #Err.	1.84	-	-	-	1.03	-	2.19	-	1.38	-	1.18	-	-	-	1.69	-	1.22	1.57	-	2.04	-	-	-	-	1.77
C→W	1.92	-	-	-	1.15	-	2.17	-	1.42	-	1.17	-	-	-	1.73	-	1.23	1.83	-	2.04	-	-	-	-	1.95
B-NLI																									
Ave. #Err.	1.43	1.12	0.98	1.01	0.96	-	-	0.99	1.42	-	-	-	-	-	-	-	1.15	1.66	-	-	1.46	1.31	1.62	1.26	1.53
C→W	1.53	1.23	1.08	1.06	0.98	-	-	1.07	1.52	-	-	-	-	-	-	-	1.27	1.72	-	-	1.60	1.39	1.64	1.35	1.67

Table 7: Average number of major errors across languages on AYA model. The first row reports the number of machine translation errors of all predictions on MT data; the second row reports the number of errors when predictions switch from correct on HT to wrong on MT.

	ar	bg	de	el	es	et	eu	fr	hi	ht	id	it	ja	ko	my	qu	ru	sw	ta	te	th	tr	ur	vi	zh
XCOPA	-	-	-	-	-	0.0	-	-	-	100	0.0	0.0	-	-	-	0.0	-	25	33.3	-	0.0	0.0	-	0.0	-50
XStoryCloze	100	-	-	-	0.0	-	-112.5	-	50.0	-	-75.0	-	-	-	12.5	-	-62.5	125	-	60.0	-	-	-	-	40.0
B-NLI	100	0.0	0.0	0.0	100	-	-	0.0	0.0	-	-	-	-	-	-	-	0.0	0.0	-	-	0.0	0.0	0.0	100	0.0

Table 8: The percentage of gap reduction between HT and MT after filtering out low quality MT examples.

ing MT.

6 Large-scale evaluation results

Having confirmed in Section 5 that our translated test sets are of a reliable quality, we now move on to analyze how the MLMs perform on them. In Figure 1, we summarize the performance of XLM-R, BLOOMz, and AYA in 196, 168, and 198 languages, categorized by their data scarcity during the pre-training of each model as explained in Section 3.3. We find that the average performance for all models is similar for high- and mid-resource languages. Yet, while still above the random baseline, there is a notable drop in performance for low-resource and unseen languages. Moreover, we find that standard deviations across low-resource languages are larger than high- and mid-resource ones. This shows that performance across low-resource languages varies a lot, making the average score less reliable. Still, performance in unseen languages is relatively high; for XNLI and PAWS-X, on average, we obtain +18% and +29% above random performance, suggesting cross-lingual knowledge transfer to the unseen languages.

6.1 Representativeness of benchmark language sets

As each dataset contains a distinct selection of languages for testing, we study to what extent each of them provides a reliable estimate for how MLMs performance will generalize to more languages.

	High	Mid	Low	Ave.
XLM-R				
PAWS-X	89.5 / 89.6	- / 87.3	- / 85.6	89.5 / 85.3
XNLI	82.2 / 81.5	79.5 / 78.2	74.4 / 71.9	80.5 / 72.4
XCOPA	71.6 / 71.8	69.6 / 68.2	66.4 / 61.7	70.3 / 69.2
XStoryCloze	85.5 / 83.1	77.2 / 78.6	74.1 / 69.9	78.9 / 77.2
BLOOMz				
B-NLI	72.8 / 73.0	70.8 / 70.2	70.9 / 68.3	72.2 / 69.8
XCOPA	76.9 / 79.0	71.6 / 67.5	63.0 / 54.3	73.7 / 62.3
XStoryCloze	86.2 / 87.9	81.1 / 72.0	79.4 / 64.5	82.8 / 71.6
AYA				
B-NLI	78.4/77.8	77.7/77.6	73.5/75.5	77.4/76.4
XCOPA	83.65/86.3	84.8/84.6	82.8/79.8	83.7/82.0
XStoryCloze	85.8/89.7	91.0/89.6	85.5/86.4	87.0/87.7

Table 9: The average performance of high, mid, and low-resource languages covered by the original dataset/the languages covered by our machine-translated datasets. All results are computed on the machine-translated data.

While we do not cover all the world's languages, we compare the averages between the languages covered by the original datasets and those covered by our much larger translated datasets that contain 198 languages.

To this end, we split the languages from the original datasets based on our resource categorization reported in Table 1, and report the average performance for each category in Table 9. Importantly, all performance scores are computed on the translated data. From the results, we observe that for high and, to some extent, mid-resource languages, average performance on both language selections

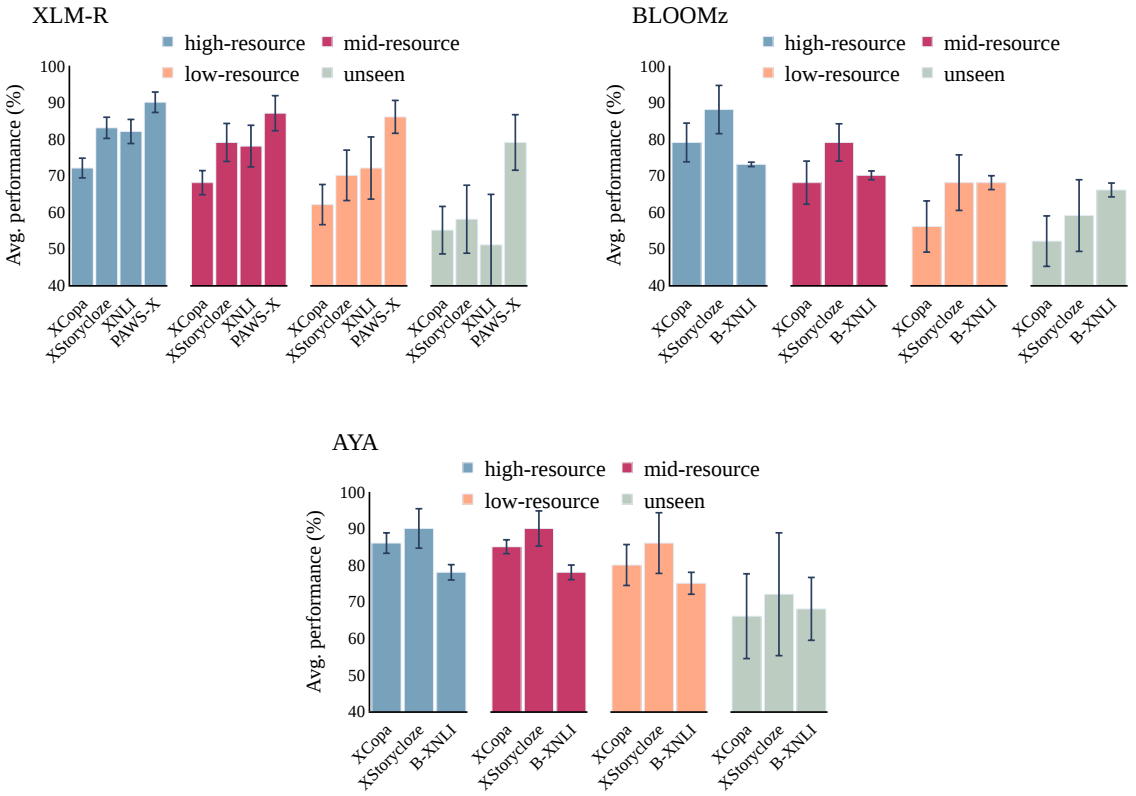


Figure 1: Average performance across test languages in a zero-shot fine-tuning setup for XLM-R and in a zero-shot prompting using BLOOMz. Results are categorized per task and data coverage during pretraining as reported in Table 1. Results across models are not directly comparable as their language categorizations differ.

is similar, making the language coverage from the original datasets sufficiently representative. Yet, for low-resource languages, we find a notable difference, which suggests that the datasets' language coverage is not representative of a wider range of low-resource languages. Specifically, across all tasks, we tend to overestimate performance, which can go up to 4.7% and 8.7% accuracy points (for XCOPA).

7 Conclusions

In this paper, we investigate the use of machine translation to create large-scale multilingual evaluation sets in scenarios where human-translated data is unavailable. Our experiments show that using SOTA MT models for evaluation yields results closely aligned with HT ones, with small average accuracy differences and high rank correlations, indicating stable relative performance across languages. Translation quality, measured by chrF++, does not strongly predict downstream performance differences, suggesting that minor variations in automatic scores are not critical for many tasks. However, LLM-as-a-judge analysis reveals that examples where models succeed on HT but fail on MT often contain more major errors, as such this method could be used to filter low-quality translations to reduce evaluation noise. Scaling evaluations to nearly 200 languages further reveals representativeness gaps in existing benchmarks. For high- and mid-resource languages, original language selections approximate broader trends well, but for low-resource languages, they overestimate performance—by up to 8.7 accuracy points in some cases. This demonstrates that MT-based evaluation can both expand coverage and uncover biases in benchmark design. Overall, our findings indicate that MT, when coupled with targeted quality control, enables broader, more representative, and more equitable multilingual evaluation, while highlighting the need to reassess how language selections in current benchmarks reflect true model capabilities.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Is abella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Jo hannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Ma teusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack W. Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll L. Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Akari Asai, Trina Chatterjee, Junjie Hu, Eunsol Choi, et al. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *2022 Findings of the Association for Computational Linguistics: EMNLP 2022*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings

of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2020. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pages 3428–3448. Association for Computational Linguistics (ACL).

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2022. Chatgpt.

ChaeHun Park, Koanho Lee, Hyesu Lim, Jaeseok Kim, Junmo Park, Yu-Jung Heo, Du-Seong Chang, and Jaegul Choo. 2024. Translation deserves better: Analyzing translation artifacts in cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5193–5221, Bangkok, Thailand. Association for Computational Linguistics.

Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel

Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. Towards multilingual llm evaluation for european languages. *CoRR*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024a. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024b. Aya model: An instruction finetuned open-access multilingual language model. *arXiv e-prints*, pages arXiv–2402.

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3544–3552.

Jiaan Wang, Fandong Meng, Yunlong Liang, Tingyi Zhang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2023. Understanding translationese in cross-lingual summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3837–3849, Singapore. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase Adversaries from Word Scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Experimental setups

For the implementation of all models, we rely on the HuggingFace Library (Wolf et al., 2019). XLM-R large, BLOOMz, and AYA-101 (AYA) have 330M, 7.1B, and 13B parameters, respectively. Moreover, we have run all the BLOOMz and AYA experiments on an NVIDIA A100-SXM4 GPU with 40GB memory, and a single NVIDIA A6000 has been used for the MT and XLM-R experiments with 48GB memory.

XLM-R fine-tuning details For the NLI task, we have fine-tuned XLM-R with a learning rate of $2e$-5, AdamW optimizer, and a batch size of 32 for 3 epochs. For the PAWS-X task, we have considered a learning rate of $2e$-6, batch size 16 with a warm-up ratio of 0.01 for 3 epochs. For XCOPA and XStoryCloze tasks, first, we train the model on the training set of Social IQa (Sap et al., 2019) and then fine-tune it on the training set of XCOPA dataset (Gordon et al., 2012). We have selected a learning rate of $3e$-6, batch size of 16 for SIQa and 8 for XCOPA, a warm-up ratio of 0.1, and fine-tune the model for 3 epochs on each dataset.

BLOOMz and AYA zero-shot prompts For zero-shot prompting, we constructed the following prompts for XNLI, XCOPA and XStorycloze respectively:

Premise: `<premise>`
Hypothesis: `<hypothesis>`
Does the premise entail the hypothesis?
Pick between yes or no.

Premise: `<premise>`
Option A: `<choice1>`
Option B: `<choice2>`
Based on the premise, which `<cause/effect>` is more likely?
Pick between options A and B.
Answer:

Consider the following story:
`<story>`
Which ending to the story is most likely?
Pick between options A and B:
A: `<story_ending1>`
B: `<story_ending2>`
Answer:

B Lessons learned for Machine Translation

We now also share a few practical lessons learned from our MT experiments using NLLB to facilitate the translation of new datasets in future work:

- NLLB tends to skip sentences when translating paragraphs. Thus, it is important to translate the sentences one by one.
- NLLB has difficulty translating short phrases/names such as names, dates, locations, etc., because it tends to hallucinate additional content. This makes it challenging to translate the answers from QA datasets such as XQuAD.
- NLLB inconsistently chooses to code-switch to the target language. For instance, when translating the sentence 'Sara is asleep', it can choose to translate it either to 'Sara ? farsi' or 'fully farsi'. This can be particularly challenging for retrieval datasets where the answer does not tend to fully match the context.
- While translation quality tends to be similar for different NLLB model sizes, at least the 3.3B version should be used when translating to languages that were low-resource, considering NLLB's pretraining data.

C Correlation between Chrf++ scores and translations

See Table 10 and Table 11 for Spearman and Pearson correlation results between Chrf++ scores and performance on human and machine translated data.

D LLM-as-a-judge

Following previous work, we use 3-shot prompting for this experiment provided in Table 12.

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
chrF++ vs. Human translated	29.6	24.4	65.7	16.4	5.7	5.7	68.1	22.7	8.1	-34.6
chrF++ vs. Machine translated	22.6	48.1	82.7	22.4	18.3	14.7	75.8	55.5	26.1	22.4

Table 10: Average Spearman rank correlation between chrf++ scores and human- (original data) and machine-translated data (ours).

	XLM-R				BLOOMz			AYA		
	XCOPA	XNLI	PAWS-X	XStoryCloze	XCOPA	XNLI	XStoryCloze	XCOPA	XNLI	XStoryCloze
chrF++ vs. Human translated	27.4	10.5	89.3	3.3	-10.4	-10.4	64.5	41.0	8.2	-18.5
chrF++ vs. Machine translated	30.3	66.2	96.5	8.8	14.5	2.9	66.7	52.1	16.3	11.3

Table 11: Average Pearson correlation between chrf++ scores and human- (original data) and machine-translated data (ours).

In-context-learning prompt:

Based on the given source, identify the major and minor errors in this translation. Note that Major errors refer to actual translation or grammatical errors, and Minor errors refer to smaller imperfections, and purely subjective opinions about the translation.

Source: I do apologise about this, we must gain permission from the account holder to discuss an order with another person, I apologise if this was done previously, however, I would not be able to discuss this with yourself without the account holder's permission.

Translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir `involvement`.

Errors:

Major: accuracy/mistranslation – `involvement`; accuracy/omission – `the account holder`

Minor: fluency/grammar – `wäre`; fluency/register – `dir`

Source: Talks have resumed in Vienna to try to revive the nuclear pact, with both sides trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.

Translation: Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemže obě partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.

Errors:

Major: accuracy/addition – `ve Vídni`; accuracy/omission – `the stop-start`

Minor: terminology/inappropriate for context – `partaje`

Source: 大众点评乌鲁木齐家居商场频道为您提供高铁居然之家地址，电话，营业时间等最新商户信息，找装修公司，就上大众点评

Translation: Urumqi Home Furnishing Store Channel provides you with the latest business information such as the address, telephone number, business hours, etc., of high-speed rail, and find a decoration company, and go to the reviews.

Errors:

Major: accuracy/addition – `of high-speed rail`; accuracy/mistranslation – `go to the reviews`

Minor: style/awkward – `etc.,`

Source: *{source}*

Translation: *{translation}*

Errors:

Table 12: The prompt used for the llm-as-a-judge experiment.

E Full results XLM-R Large

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	71.05	fao_Latn	87.4	lij_Latn	86.2	slv_Latn	89.75
ace_Latn	85.05	fij_Latn	75.45	lim_Latn	87.55	smo_Latn	78.6
acm_Arab	88.5	fin_Latn	86.2	lin_Latn	77.25	sna_Latn	78.6
acq_Arab	87.85	fon_Latn	69.6	lit_Latn	89.2	snd_Arab	86.95
aeb_Arab	87.2	fra_Latn	91.7	lmo_Latn	86.9	som_Latn	86.15
afr_Latn	92.1	fur_Latn	86.9	ltg_Latn	81.4	sot_Latn	80.45
ajp_Arab	88.45	fuv_Latn	74.3	ltz_Latn	85.6	spa_Latn	92.2
aka_Latn	75.45	gaz_Latn	77.75	lua_Latn	75.65	srd_Latn	87.25
als_Latn	92.25	gla_Latn	88.15	lug_Latn	76.95	srp_Cyrl	90.65
amh_Ethi	82.85	gle_Latn	88.1	luo_Latn	73.95	ssw_Latn	80.15
apc_Arab	87.1	glg_Latn	91.75	lus_Latn	77.3	sun_Latn	89.6
arb_Arab	88.7	grn_Latn	79.75	lvs_Latn	87.5	swe_Latn	91.1
ars_Arab	89.15	guj_Gujr	87.75	mag_Deva	88.35	swh_Latn	89.65
ary_Arab	87.25	hat_Latn	86.5	mai_Deva	84.9	szl_Latn	87.45
arz_Arab	89.25	hau_Latn	86.15	mal_Mlym	82.9	tam_Taml	84.35
asm_Beng	83.2	heb_Hebr	89.05	mar_Deva	83.25	taq_Latn	73.25
ast_Latn	86.95	hin_Deva	88.4	min_Latn	88.4	tat_Cyrl	79.5
awa_Deva	84.95	hne_Deva	86.45	mkd_Cyrl	90.75	tel_Telu	85.65
ayr_Latn	71.45	hrv_Latn	90.6	mlt_Latn	82.65	tgk_Cyrl	77.65
azb_Arab	72.1	hun_Latn	88.65	mni_Beng	74.2	tgl_Latn	92.6
azj_Latn	86.0	hye_Armn	87.55	mos_Latn	69.95	tha_Thai	87.7
bak_Cyrl	78.45	ibo_Latn	77.8	mri_Latn	79.75	tir_Ethi	77.85
bam_Latn	78.35	ilo_Latn	84.4	mya_Mymr	81.45	tpi_Latn	74.4
ban_Latn	84.25	ind_Latn	92.6	nld_Latn	90.55	tsn_Latn	80.5
bel_Cyrl	88.5	isl_Latn	88.15	nno_Latn	84.15	tso_Latn	76.8
bem_Latn	74.0	ita_Latn	92.25	nob_Latn	92.7	tuk_Latn	74.25
ben_Beng	86.8	jav_Latn	89.4	npi_Deva	86.45	tum_Latn	73.0
bho_Deva	84.55	jpn_Jpan	83.25	nso_Latn	82.5	tur_Latn	85.8
bjn_Arab	70.4	kab_Latn	75.9	nus_Latn	71.5	twi_Latn	75.0
bjn_Latn	89.7	kac_Latn	72.05	nya_Latn	78.2	uig_Arab	80.25
bos_Latn	91.25	kam_Latn	64.25	oci_Latn	90.7	ukr_Cyrl	88.8
bug_Latn	81.25	kan_Knda	87.0	pag_Latn	82.15	umb_Latn	65.9
bul_Cyrl	90.25	kas_Arab	75.2	pan_Guru	87.4	urd_Arab	88.5
cat_Latn	92.1	kas_Deva	75.65	pap_Latn	87.45	uzn_Latn	84.0
ceb_Latn	87.75	kat_Geor	79.1	pbt_Arab	88.0	vec_Latn	89.95
ces_Latn	91.25	kaz_Cyrl	83.55	pes_Arab	87.85	vie_Latn	90.85
cjk_Latn	72.65	kbp_Latn	71.2	plt_Latn	87.2	war_Latn	85.1
ckb_Arab	76.5	kea_Latn	86.65	pol_Latn	90.2	wol_Latn	73.95
crh_Latn	83.8	khk_Cyrl	83.0	por_Latn	91.9	xho_Latn	85.55
cym_Latn	89.2	khm_Khmr	86.35	prs_Arab	89.2	ydd_Hebr	89.2
dan_Latn	92.25	kik_Latn	74.1	quy_Latn	68.7	yor_Latn	73.95
deu_Latn	90.6	kin_Latn	75.55	ron_Latn	91.2	yue_Hant	81.8
dik_Latn	72.85	kir_Cyrl	80.75	run_Latn	74.1	zho_Hans	82.4
dyu_Latn	69.45	kmb_Latn	67.7	rus_Cyrl	90.1	zho_Hant	67.15
ell_Grek	89.7	kmr_Latn	85.45	sag_Latn	74.05	zsm_Latn	92.6
eng_Latn	94.0	knc_Arab	77.7	san_Deva	75.35	zul_Latn	85.95
epo_Latn	91.3	knc_Latn	71.0	scn_Latn	87.8		
est_Latn	87.55	kon_Latn	75.0	shn_Mymr	68.1		
eus_Latn	78.65	kor_Hang	85.3	sin_Sinh	84.3		
ewe_Latn	75.9	lao_Laoo	89.25	slk_Latn	91.1		

Figure 2: The accuracy score of XLM-R model on PAWS-X task across 196 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	36.89	fao_Latn	58.26	lij_Latn	63.15	slv_Latn	80.4
ace_Latn	52.93	fij_Latn	41.34	lim_Latn	67.6	smo_Latn	40.68
acm_Arab	75.71	fin_Latn	80.02	lin_Latn	40.52	sna_Latn	40.82
acq_Arab	78.08	fon_Latn	40.2	lit_Latn	80.74	snd_Arab	77.05
aeb_Arab	65.63	fra_Latn	82.87	lmo_Latn	65.53	som_Latn	67.68
afr_Latn	83.53	fur_Latn	61.62	ltg_Latn	56.85	sot_Latn	40.74
ajp_Arab	75.63	fuv_Latn	39.9	ltz_Latn	50.56	spa_Latn	84.93
aka_Latn	40.14	gaz_Latn	56.53	lua_Latn	39.92	srd_Latn	65.19
als_Latn	81.22	gla_Latn	70.76	lug_Latn	42.32	srp_Cyrl	82.69
amh_Ethi	69.1	gle_Latn	75.05	luo_Latn	39.5	ssw_Latn	44.31
apc_Arab	72.87	glg_Latn	84.35	lus_Latn	42.38	sun_Latn	74.03
arb_Arab	80.16	grn_Latn	43.49	lvs_Latn	77.62	swe_Latn	82.99
ars_Arab	76.91	guj_Gujr	77.52	mag_Deva	70.74	swh_Latn	74.39
ary_Arab	69.24	hat_Latn	44.91	mai_Deva	67.05	szl_Latn	71.84
arz_Arab	77.92	hau_Latn	68.96	mal_Mlym	76.45	tam_Taml	76.37
asm_Beng	72.57	heb_Hebr	82.48	mar_Deva	78.26	taq_Latn	38.38
ast_Latn	70.66	hin_Deva	80.36	min_Latn	62.22	tat_Cyrl	48.24
awa_Deva	62.28	hne_Deva	69.6	mkd_Cyrl	82.02	tel_Telu	75.65
ayr_Latn	40.36	hrv_Latn	81.86	mlt_Latn	43.27	tgk_Cyrl	39.42
azb_Arab	40.42	hun_Latn	81.4	mni_Beng	36.39	tgl_Latn	79.34
azj_Latn	78.54	hye_Armn	78.28	mos_Latn	38.98	tha_Thai	76.93
bak_Cyrl	45.53	ibo_Latn	41.32	mri_Latn	40.12	tir_Ethi	45.55
bam_Latn	40.82	ilo_Latn	43.03	mya_Mymr	72.02	tpi_Latn	48.16
ban_Latn	49.54	ind_Latn	83.65	nld_Latn	83.09	tsn_Latn	41.36
bel_Cyrl	81.58	isl_Latn	77.35	nno_Latn	74.91	tso_Latn	41.4
bem_Latn	40.18	ita_Latn	84.11	nob_Latn	84.89	tuk_Latn	49.22
ben_Beng	77.86	jav_Latn	75.63	npi_Deva	74.71	tum_Latn	40.16
bho_Deva	71.76	jpn_Jpan	73.91	nso_Latn	41.28	tur_Latn	79.98
bjn_Arab	36.25	kab_Latn	37.39	nus_Latn	38.28	twi_Latn	40.3
bjn_Latn	65.79	kac_Latn	39.38	nya_Latn	42.04	uig_Arab	63.71
bos_Latn	82.36	kam_Latn	36.63	oci_Latn	71.14	ukr_Cyrl	81.64
bug_Latn	43.69	kan_Knda	77.47	pag_Latn	41.82	umb_Latn	37.09
bul_Cyrl	82.99	kas_Arab	48.54	pan_Guru	77.03	urd_Arab	78.12
cat_Latn	83.11	kas_Deva	50.78	pap_Latn	63.19	uzn_Latn	77.62
ceb_Latn	57.27	kat_Geor	56.23	pbt_Arab	78.32	vec_Latn	75.35
ces_Latn	82.26	kaz_Cyrl	76.25	pes_Arab	79.18	vie_Latn	80.72
cjk_Latn	40.02	kbp_Latn	37.37	plt_Latn	69.44	war_Latn	48.84
ckb_Arab	40.26	kea_Latn	51.64	pol_Latn	81.88	wol_Latn	40.58
crh_Latn	64.55	khk_Cyrl	74.59	por_Latn	84.23	xho_Latn	60.0
cym_Latn	79.04	khm_Khmr	73.87	prs_Arab	80.0	ydd_Hebr	74.81
dan_Latn	83.07	kik_Latn	38.54	quy_Latn	38.08	yor_Latn	38.88
deu_Latn	82.79	kin_Latn	39.68	ron_Latn	83.27	yue_Hant	70.16
dik_Latn	39.84	kir_Cyrl	73.99	run_Latn	40.42	zho_Hans	72.42
dyu_Latn	36.59	kmb_Latn	37.58	rus_Cyrl	81.76	zho_Hant	60.08
ell_Grek	82.12	kmr_Latn	72.59	sag_Latn	42.95	zsm_Latn	82.34
eng_Latn	87.25	knc_Arab	59.5	san_Deva	68.16	zul_Latn	60.0
epo_Latn	80.56	knc_Latn	39.34	scn_Latn	65.53		
est_Latn	80.92	kon_Latn	40.96	shn_Mymr	37.07		
eus_Latn	66.57	kor_Hang	77.86	sin_Sinh	75.39		
ewe_Latn	41.32	lao_Laoo	77.94	slk_Latn	82.1		

Figure 3: The accuracy score of XLM-R model on XNLI task across 196 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	49.44	fao_Latn	65.12	lij_Latn	59.7	slv_Latn	82.06
ace_Latn	58.24	fij_Latn	53.61	lim_Latn	72.73	smo_Latn	50.83
acm_Arab	72.27	fin_Latn	85.04	lin_Latn	49.77	sna_Latn	52.15
acq_Arab	79.29	fon_Latn	50.76	lit_Latn	78.56	snd_Arab	75.25
aeb_Arab	65.78	fra_Latn	81.47	lmo_Latn	63.86	som_Latn	64.2
afr_Latn	79.95	fur_Latn	60.89	ltg_Latn	60.42	sot_Latn	52.08
ajp_Arab	77.5	fuv_Latn	50.3	ltz_Latn	58.9	spa_Latn	84.65
aka_Latn	51.82	gaz_Latn	57.58	lua_Latn	52.42	srd_Latn	61.09
als_Latn	79.09	gla_Latn	63.73	lug_Latn	50.5	srp_Cyrl	78.76
amh_Ethi	66.05	gle_Latn	71.34	luo_Latn	53.61	ssw_Latn	57.05
apc_Arab	76.51	glg_Latn	84.18	lus_Latn	54.73	sun_Latn	72.87
arb_Arab	80.28	grn_Latn	54.8	lvs_Latn	77.1	swe_Latn	86.37
ars_Arab	80.28	guj_Gujr	73.46	mag_Deva	67.97	swh_Latn	75.71
ary_Arab	73.06	hat_Latn	54.53	mai_Deva	64.2	szl_Latn	70.35
arz_Arab	77.7	hau_Latn	68.03	mal_Mlym	77.9	tam_Taml	76.04
asm_Beng	64.2	heb_Hebr	82.79	mar_Deva	77.75	taq_Latn	50.96
ast_Latn	74.39	hin_Deva	78.29	min_Latn	62.94	tat_Cyrl	56.85
awa_Deva	64.2	hne_Deva	66.91	mkd_Cyrl	82.06	tel_Telu	76.31
ayr_Latn	51.29	hrv_Latn	83.52	mlt_Latn	50.43	tgk_Cyrl	52.55
azb_Arab	54.33	hun_Latn	82.26	mni_Beng	51.56	tgl_Latn	73.53
azj_Latn	80.15	hye_Armn	76.17	mos_Latn	50.36	tha_Thai	85.51
bak_Cyrl	58.5	ibo_Latn	51.95	mri_Latn	51.09	tir_Ethi	54.33
bam_Latn	51.95	ilo_Latn	54.47	mya_Mymr	70.2	tpi_Latn	49.9
ban_Latn	58.04	ind_Latn	88.75	nld_Latn	84.71	tsn_Latn	52.61
bel_Cyrl	78.49	isl_Latn	76.24	nno_Latn	79.95	tso_Latn	52.81
bem_Latn	54.2	ita_Latn	82.33	nob_Latn	85.9	tuk_Latn	59.96
ben_Beng	75.25	jav_Latn	73.46	npi_Deva	74.92	tum_Latn	51.75
bho_Deva	70.62	jpn_Jpan	79.95	nso_Latn	52.15	tur_Latn	84.25
bjn_Arab	50.63	kab_Latn	48.97	nus_Latn	52.68	twi_Latn	51.09
bjn_Latn	69.23	kac_Latn	52.35	nya_Latn	50.36	uig_Arab	70.22
bos_Latn	81.67	kam_Latn	50.69	oci_Latn	71.87	ukr_Cyrl	82.26
bug_Latn	53.34	kan_Knda	74.72	pag_Latn	53.28	umb_Latn	51.75
bul_Cyrl	82.46	kas_Arab	54.53	pan_Guru	70.46	urd_Arab	76.37
cat_Latn	81.4	kas_Deva	64.2	pap_Latn	59.3	uzn_Latn	73.4
ceb_Latn	60.36	kat_Geor	56.98	pbt_Arab	72.14	vec_Latn	74.52
ces_Latn	83.06	kaz_Cyrl	76.7	pes_Arab	79.75	vie_Latn	82.53
cjk_Latn	49.83	kbp_Latn	52.08	plt_Latn	64.99	war_Latn	58.44
ckb_Arab	51.69	kea_Latn	58.44	pol_Latn	83.26	wol_Latn	50.23
crh_Latn	69.82	khk_Cyrl	75.84	por_Latn	84.32	xho_Latn	56.98
cym_Latn	71.74	khm_Khmr	72.07	prs_Arab	81.14	ydd_Hebr	63.14
dan_Latn	85.9	kik_Latn	50.83	quy_Latn	47.39	yor_Latn	51.62
deu_Latn	82.53	kin_Latn	49.5	ron_Latn	80.94	yue_Hant	79.95
dik_Latn	52.68	kir_Cyrl	75.78	run_Latn	47.92	zho_Hans	86.9
dyu_Latn	50.43	kmb_Latn	50.63	rus_Cyrl	83.45	zho_Hant	79.62
ell_Grek	76.84	kmr_Latn	69.03	sag_Latn	52.28	zsm_Latn	87.29
eng_Latn	88.82	knc_Arab	58.9	san_Deva	65.59	zul_Latn	58.7
epo_Latn	78.89	knc_Latn	51.42	scn_Latn	66.91		
est_Latn	82.73	kon_Latn	53.61	shn_Mymr	46.39		
eus_Latn	76.37	kor_Hang	78.42	sin_Sinh	77.43		
ewe_Latn	52.68	lao_Laoo	76.37	slk_Latn	83.85		

Figure 4: The accuracy score of XLM-R model on XStoryCloze task across 196 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	50.0	fao_Latn	61.6	lij_Latn	62.0	slv_Latn	72.6
ace_Latn	54.8	fij_Latn	51.6	lim_Latn	64.2	smo_Latn	44.4
acm_Arab	67.0	fin_Latn	72.0	lin_Latn	51.4	sna_Latn	49.4
acq_Arab	66.6	fon_Latn	48.8	lit_Latn	69.6	snd_Arab	64.2
aeb_Arab	60.0	fra_Latn	73.6	lmo_Latn	60.6	som_Latn	57.0
afr_Latn	71.0	fur_Latn	56.2	ltg_Latn	59.6	sot_Latn	49.2
ajp_Arab	68.4	fuv_Latn	49.4	ltz_Latn	57.0	spa_Latn	73.4
aka_Latn	50.4	gaz_Latn	50.0	lua_Latn	48.8	srd_Latn	59.4
als_Latn	67.2	gla_Latn	57.6	lug_Latn	50.4	srp_Cyrl	66.8
amh_Ethi	61.8	gle_Latn	61.0	luo_Latn	50.0	ssw_Latn	52.6
apc_Arab	66.4	glg_Latn	72.4	lus_Latn	50.8	sun_Latn	60.4
arb_Arab	71.2	grn_Latn	48.8	lvs_Latn	68.6	swe_Latn	74.6
ars_Arab	69.6	guj_Gujr	64.6	mag_Deva	60.2	swh_Latn	66.4
ary_Arab	64.8	hat_Latn	50.4	mai_Deva	58.6	szl_Latn	63.4
arz_Arab	70.8	hau_Latn	54.4	mal_Mlym	66.6	tam_Taml	71.8
asm_Beng	62.2	heb_Hebr	67.0	mar_Deva	67.2	taq_Latn	48.4
ast_Latn	65.4	hin_Deva	66.4	min_Latn	53.8	tat_Cyrl	47.0
awa_Deva	64.2	hne_Deva	62.2	mkd_Cyrl	69.6	tel_Telu	65.8
ayr_Latn	48.2	hrv_Latn	71.0	mlt_Latn	51.4	tgk_Cyrl	50.2
azb_Arab	54.2	hun_Latn	68.6	mni_Beng	52.4	tgl_Latn	67.4
azj_Latn	67.2	hye_Armn	67.8	mos_Latn	48.4	tha_Thai	69.4
bak_Cyrl	50.4	ibo_Latn	49.0	mri_Latn	49.6	tir_Ethi	50.2
bam_Latn	51.4	ilo_Latn	50.8	mya_Mymr	65.4	tpi_Latn	50.0
ban_Latn	57.0	ind_Latn	74.6	nld_Latn	71.4	tsn_Latn	50.6
bel_Cyrl	64.8	isl_Latn	65.6	nno_Latn	70.6	tso_Latn	51.4
bem_Latn	53.0	ita_Latn	74.0	nob_Latn	70.6	tuk_Latn	51.6
ben_Beng	67.0	jav_Latn	59.8	npi_Deva	64.13	tum_Latn	50.6
bho_Deva	66.4	jpn_Jpan	72.4	nso_Latn	55.4	tur_Latn	68.6
bjn_Arab	49.8	kab_Latn	51.4	nus_Latn	49.0	twi_Latn	50.2
bjn_Latn	62.2	kac_Latn	50.0	nya_Latn	51.0	uig_Arab	60.4
bos_Latn	70.4	kam_Latn	52.4	oci_Latn	64.0	ukr_Cyrl	68.4
bug_Latn	51.2	kan_Knda	67.6	pag_Latn	51.6	umb_Latn	49.2
bul_Cyrl	70.8	kas_Arab	50.4	pan_Guru	61.8	urd_Arab	69.0
cat_Latn	72.0	kas_Deva	56.2	pap_Latn	55.6	uzn_Latn	64.2
ceb_Latn	59.0	kat_Geor	55.8	pbt_Arab	66.8	vec_Latn	62.4
ces_Latn	70.4	kaz_Cyrl	65.8	pes_Arab	66.4	vie_Latn	69.2
cjk_Latn	50.2	kbp_Latn	48.2	plt_Latn	58.2	war_Latn	54.0
ckb_Arab	47.2	kea_Latn	52.0	pol_Latn	70.8	wol_Latn	51.2
crh_Latn	60.2	khk_Cyrl	66.2	por_Latn	72.8	xho_Latn	52.8
cym_Latn	63.6	khm_Khmr	68.8	prs_Arab	69.6	ydd_Hebr	62.6
dan_Latn	74.8	kik_Latn	50.6	quy_Latn	50.0	yor_Latn	45.2
deu_Latn	74.2	kin_Latn	51.4	ron_Latn	73.8	yue_Hant	68.6
dik_Latn	49.2	kir_Cyrl	65.4	run_Latn	52.4	zho_Hans	70.6
dyu_Latn	52.0	kmb_Latn	49.6	rus_Cyrl	69.6	zho_Hant	72.0
ell_Grek	73.2	kmr_Latn	62.4	sag_Latn	52.2	zsm_Latn	73.0
eng_Latn	78.6	knc_Arab	54.4	san_Deva	58.8	zul_Latn	49.6
epo_Latn	67.0	knc_Latn	48.8	scn_Latn	63.8		
est_Latn	68.4	kon_Latn	51.0	shn_Mymr	50.0		
eus_Latn	59.6	kor_Hang	70.8	sin_Sinh	66.0		
ewe_Latn	51.8	lao_Laoo	67.2	slk_Latn	69.0		

Figure 5: The accuracy score of XLM-R model on XCOPA task across 196 languages.

F Full results BLOOMz

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	64.35	fur_Latn	65.45	mar_Deva	71.68	tum_Latn	66.49
ace_Latn	66.27	fuv_Latn	66.71	min_Latn	67.09	tur_Latn	61.68
acm_Arab	68.54	gaz_Latn	66.39	mlt_Latn	64.69	twi_Latn	66.73
acq_Arab	68.82	gla_Latn	66.35	mni_Beng	62.16	uig_Arab	59.8
aeb_Arab	67.15	gle_Latn	66.51	mos_Latn	66.75	umb_Latn	65.85
afr_Latn	65.87	glg_Latn	71.28	mri_Latn	66.83	urd_Arab	69.76
ajp_Arab	68.26	grn_Latn	66.33	nld_Latn	65.89	uzn_Latn	65.33
aka_Latn	67.03	guj_Gujr	70.44	nno_Latn	66.63	vec_Latn	66.59
als_Latn	62.89	hat_Latn	65.37	nob_Latn	65.95	vie_Latn	73.01
apc_Arab	67.66	hau_Latn	65.77	npi_Deva	70.86	war_Latn	64.97
arb_Arab	72.4	hin_Deva	71.9	nso_Latn	68.76	wol_Latn	65.43
ars_Arab	69.16	hne_Deva	65.85	nus_Latn	66.01	xho_Latn	66.83
ary_Arab	68.34	hrv_Latn	66.85	nya_Latn	67.86	yor_Latn	68.04
arz_Arab	68.82	hun_Latn	67.15	oci_Latn	65.05	yue_Hant	70.38
asm_Beng	68.16	ibo_Latn	68.32	pag_Latn	66.71	zho_Hans	72.18
ast_Latn	67.84	ilo_Latn	66.31	pap_Latn	65.53	zho_Hant	68.14
awa_Deva	67.33	ind_Latn	72.26	pbt_Arab	64.07	zsm_Latn	71.76
ayr_Latn	66.33	isl_Latn	66.43	pes_Arab	62.28	zul_Latn	67.01
azb_Arab	63.91	ita_Latn	70.36	plt_Latn	66.33		
azj_Latn	59.6	jav_Latn	64.27	pol_Latn	65.81		
bam_Latn	65.55	kab_Latn	66.01	por_Latn	74.09		
ban_Latn	65.97	kac_Latn	66.19	prs_Arab	62.0		
bem_Latn	67.09	kam_Latn	66.25	quy_Latn	66.73		
ben_Beng	70.46	kan_Knda	71.04	ron_Latn	65.17		
bho_Deva	68.62	kas_Arab	63.47	run_Latn	67.6		
bjn_Arab	64.57	kas_Deva	65.63	sag_Latn	66.79		
bjn_Latn	67.82	kbp_Latn	63.23	san_Deva	65.47		
bos_Latn	67.37	kea_Latn	64.15	scn_Latn	63.75		
bug_Latn	64.95	kik_Latn	65.95	slk_Latn	66.69		
cat_Latn	73.47	kin_Latn	68.02	slv_Latn	67.19		
ceb_Latn	66.05	kmb_Latn	65.99	smo_Latn	66.07		
ces_Latn	66.07	kmr_Latn	65.01	sna_Latn	67.92		
cjk_Latn	66.01	knc_Arab	65.87	snd_Arab	64.73		
ckb_Arab	61.42	knc_Latn	65.13	som_Latn	66.15		
crh_Latn	62.24	kon_Latn	66.91	sot_Latn	67.01		
cym_Latn	65.81	lij_Latn	64.27	spa_Latn	73.59		
dan_Latn	66.77	lim_Latn	64.89	srd_Latn	65.99		
deu_Latn	68.34	lin_Latn	68.62	ssw_Latn	66.37		
dik_Latn	64.59	lit_Latn	66.27	sun_Latn	64.77		
dyu_Latn	66.01	lmo_Latn	64.47	swe_Latn	66.27		
eng_Latn	72.63	ltg_Latn	65.83	swh_Latn	70.94		
epo_Latn	63.57	ltz_Latn	64.37	szl_Latn	65.13		
est_Latn	66.25	lua_Latn	67.33	tam_Taml	70.66		
eus_Latn	68.86	lug_Latn	67.52	taq_Latn	64.89		
ewe_Latn	66.53	luo_Latn	66.33	tel_Telu	70.64		
fao_Latn	67.01	lus_Latn	65.73	tgl_Latn	65.79		
fij_Latn	67.03	lvs_Latn	64.05	tpi_Latn	66.89		
fin_Latn	65.85	mag_Deva	67.82	tsn_Latn	67.78		
fon_Latn	66.57	mai_Deva	67.58	tso_Latn	67.27		
fra_Latn	73.13	mal_Mlym	71.36	tuk_Latn	62.34		

Figure 6: The accuracy score of BLOOMz model on B-NLI task across 168 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	11.18	fur_Latn	67.57	mar_Deva	73.2	tum_Latn	64.06
ace_Latn	60.29	fuv_Latn	51.69	min_Latn	64.46	tur_Latn	50.89
acm_Arab	81.07	gaz_Latn	54.07	mlt_Latn	56.12	twi_Latn	57.45
acq_Arab	85.31	gla_Latn	50.56	mni_Beng	53.67	uig_Arab	45.47
aeb_Arab	65.25	gle_Latn	54.27	mos_Latn	50.43	umb_Latn	48.25
afr_Latn	58.97	glg_Latn	86.83	mri_Latn	51.89	urd_Arab	69.49
ajp_Arab	83.85	grn_Latn	56.25	nld_Latn	64.06	uzn_Latn	51.29
aka_Latn	56.85	guj_Gujr	79.75	nno_Latn	60.62	vec_Latn	75.65
als_Latn	53.14	hat_Latn	60.49	nob_Latn	61.48	vie_Latn	89.81
apc_Arab	80.54	hau_Latn	53.87	npi_Deva	70.75	war_Latn	55.53
arb_Arab	88.02	hin_Deva	84.45	nso_Latn	65.65	wol_Latn	55.39
ars_Arab	86.1	hne_Deva	30.18	nus_Latn	49.17	xho_Latn	65.65
ary_Arab	73.86	hrv_Latn	55.86	nya_Latn	68.3	yor_Latn	71.54
arz_Arab	84.05	hun_Latn	51.16	oci_Latn	82.26	yue_Hant	42.22
asm_Beng	79.09	ibo_Latn	66.84	pag_Latn	53.47	zho_Hans	69.62
ast_Latn	82.26	ilo_Latn	56.06	pap_Latn	64.33	zho_Hant	31.5
awa_Deva	30.77	ind_Latn	89.54	pbt_Arab	21.91	zsm_Latn	88.48
ayr_Latn	56.85	isl_Latn	51.89	pes_Arab	39.44	zul_Latn	66.51
azb_Arab	1.13	ita_Latn	82.73	plt_Latn	55.26		
azj_Latn	52.08	jav_Latn	64.39	pol_Latn	56.59		
bam_Latn	56.92	kab_Latn	53.01	por_Latn	90.14		
ban_Latn	57.58	kac_Latn	53.87	prs_Arab	45.8		
bem_Latn	54.53	kam_Latn	52.15	quy_Latn	54.73		
ben_Beng	84.32	kan_Knda	76.7	ron_Latn	62.94		
bho_Deva	31.17	kas_Arab	2.65	run_Latn	61.02		
bjn_Arab	20.58	kas_Deva	32.3	sag_Latn	53.54		
bjn_Latn	72.14	kbp_Latn	49.31	san_Deva	34.81		
bos_Latn	57.38	kea_Latn	62.34	scn_Latn	66.71		
bug_Latn	53.74	kik_Latn	57.45	slk_Latn	55.33		
cat_Latn	89.94	kin_Latn	62.94	slv_Latn	54.6		
ceb_Latn	55.79	kmb_Latn	49.7	smo_Latn	52.55		
ces_Latn	56.52	kmr_Latn	52.68	sna_Latn	66.12		
cjk_Latn	52.15	knc_Arab	62.08	snd_Arab	6.42		
ckb_Arab	31.7	knc_Latn	53.01	som_Latn	51.95		
crh_Latn	51.62	kon_Latn	53.67	sot_Latn	65.19		
cym_Latn	51.95	lij_Latn	64.39	spa_Latn	91.0		
dan_Latn	62.01	lim_Latn	60.23	srd_Latn	63.4		
deu_Latn	71.94	lin_Latn	63.8	ssw_Latn	59.43		
dik_Latn	53.01	lit_Latn	54.73	sun_Latn	63.4		
dyu_Latn	54.4	lmo_Latn	70.55	swe_Latn	60.82		
eng_Latn	93.12	ltg_Latn	53.54	swh_Latn	79.09		
epo_Latn	64.99	ltz_Latn	58.57	szl_Latn	54.14		
est_Latn	51.29	lua_Latn	50.36	tam_Taml	76.64		
eus_Latn	77.83	lug_Latn	62.01	taq_Latn	53.01		
ewe_Latn	53.41	luo_Latn	53.34	tel_Telu	72.73		
fao_Latn	54.86	lus_Latn	54.86	tgl_Latn	54.8		
fij_Latn	52.02	lvs_Latn	53.08	tpi_Latn	56.78		
fin_Latn	53.28	mag_Deva	36.6	tsn_Latn	64.99		
fon_Latn	38.19	mai_Deva	41.69	tso_Latn	64.92		
fra_Latn	89.94	mal_Mlym	80.08	tuk_Latn	50.3		

Figure 7: The accuracy score of BLOOMz model on XStoryCloze task across 168 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	64.35	fur_Latn	65.45	mar_Deva	71.68	tum_Latn	66.49
ace_Latn	66.27	fuv_Latn	66.71	min_Latn	67.09	tur_Latn	61.68
acm_Arab	68.54	gaz_Latn	66.39	mlt_Latn	64.69	twi_Latn	66.73
acq_Arab	68.82	gla_Latn	66.35	mni_Beng	62.16	uig_Arab	59.8
aeb_Arab	67.15	gle_Latn	66.51	mos_Latn	66.75	umb_Latn	65.85
afr_Latn	65.87	glg_Latn	71.28	mri_Latn	66.83	urd_Arab	69.76
ajp_Arab	68.26	grn_Latn	66.33	nld_Latn	65.89	uzn_Latn	65.33
aka_Latn	67.03	guj_Gujr	70.44	nno_Latn	66.63	vec_Latn	66.59
als_Latn	62.89	hat_Latn	65.37	nob_Latn	65.95	vie_Latn	73.01
apc_Arab	67.66	hau_Latn	65.77	npi_Deva	70.86	war_Latn	64.97
arb_Arab	72.4	hin_Deva	71.9	nso_Latn	68.76	wol_Latn	65.43
ars_Arab	69.16	hne_Deva	65.85	nus_Latn	66.01	xho_Latn	66.83
ary_Arab	68.34	hrv_Latn	66.85	nya_Latn	67.86	yor_Latn	68.04
arz_Arab	68.82	hun_Latn	67.15	oci_Latn	65.05	yue_Hant	70.38
asm_Beng	68.16	ibo_Latn	68.32	pag_Latn	66.71	zho_Hans	72.18
ast_Latn	67.84	ilo_Latn	66.31	pap_Latn	65.53	zho_Hant	68.14
awa_Deva	67.33	ind_Latn	72.26	pbt_Arab	64.07	zsm_Latn	71.76
ayr_Latn	66.33	isl_Latn	66.43	pes_Arab	62.28	zul_Latn	67.01
azb_Arab	63.91	ita_Latn	70.36	plt_Latn	66.33		
azj_Latn	59.6	jav_Latn	64.27	pol_Latn	65.81		
bam_Latn	65.55	kab_Latn	66.01	por_Latn	74.09		
ban_Latn	65.97	kac_Latn	66.19	prs_Arab	62.0		
bem_Latn	67.09	kam_Latn	66.25	quy_Latn	66.73		
ben_Beng	70.46	kan_Knda	71.04	ron_Latn	65.17		
bho_Deva	68.62	kas_Arab	63.47	run_Latn	67.6		
bjn_Arab	64.57	kas_Deva	65.63	sag_Latn	66.79		
bjn_Latn	67.82	kbp_Latn	63.23	san_Deva	65.47		
bos_Latn	67.37	kea_Latn	64.15	scn_Latn	63.75		
bug_Latn	64.95	kik_Latn	65.95	slk_Latn	66.69		
cat_Latn	73.47	kin_Latn	68.02	slv_Latn	67.19		
ceb_Latn	66.05	kmb_Latn	65.99	smo_Latn	66.07		
ces_Latn	66.07	kmr_Latn	65.01	sna_Latn	67.92		
cjk_Latn	66.01	knc_Arab	65.87	snd_Arab	64.73		
ckb_Arab	61.42	knc_Latn	65.13	som_Latn	66.15		
crh_Latn	62.24	kon_Latn	66.91	sot_Latn	67.01		
cym_Latn	65.81	lij_Latn	64.27	spa_Latn	73.59		
dan_Latn	66.77	lim_Latn	64.89	srd_Latn	65.99		
deu_Latn	68.34	lin_Latn	68.62	ssw_Latn	66.37		
dik_Latn	64.59	lit_Latn	66.27	sun_Latn	64.77		
dyu_Latn	66.01	lmo_Latn	64.47	swe_Latn	66.27		
eng_Latn	72.63	ltg_Latn	65.83	swh_Latn	70.94		
epo_Latn	63.57	ltz_Latn	64.37	szl_Latn	65.13		
est_Latn	66.25	lua_Latn	67.33	tam_Taml	70.66		
eus_Latn	68.86	lug_Latn	67.52	taq_Latn	64.89		
ewe_Latn	66.53	luo_Latn	66.33	tel_Telu	70.64		
fao_Latn	67.01	lus_Latn	65.73	tgl_Latn	65.79		
fij_Latn	67.03	lvs_Latn	64.05	tpi_Latn	66.89		
fin_Latn	65.85	mag_Deva	67.82	tsn_Latn	67.78		
fon_Latn	66.57	mai_Deva	67.58	tso_Latn	67.27		
fra_Latn	73.13	mal_Mlym	71.36	tuk_Latn	62.34		

Figure 8: The accuracy score of BLOOMz model on XCOPA task across 168 languages.

G Full Results for AYA

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	55.0	fao_Latn	88.0	lij_Latn	84.0	slv_Latn	92.0
ace_Latn	81.0	fij_Latn	73.0	lim_Latn	89.0	smo_Latn	89.0
acm_Arab	88.0	fin_Latn	71.0	lin_Latn	72.0	sna_Latn	85.0
acq_Arab	90.0	fon_Latn	41.0	lit_Latn	90.0	snd_Arab	87.0
aeb_Arab	78.0	fra_Latn	93.0	lmo_Latn	86.0	som_Latn	83.0
afr_Latn	94.0	fur_Latn	86.0	ltg_Latn	77.0	sot_Latn	80.0
ajp_Arab	87.0	fuv_Latn	51.0	ltz_Latn	90.0	spa_Latn	94.0
aka_Latn	73.0	gaz_Latn	73.0	lua_Latn	67.0	srd_Latn	88.0
als_Latn	92.0	gla_Latn	86.0	lug_Latn	75.0	srp_Cyrl	93.0
amh_Ethi	80.0	gle_Latn	90.0	luo_Latn	51.0	ssw_Latn	77.0
apc_Arab	86.0	glg_Latn	90.0	lus_Latn	70.0	sun_Latn	90.0
arb_Arab	92.0	grn_Latn	55.0	lvs_Latn	89.0	swe_Latn	91.0
ars_Arab	90.0	guj_Gujr	89.0	mag_Deva	89.0	swh_Latn	89.0
ary_Arab	85.0	hat_Latn	91.0	mai_Deva	88.0	szl_Latn	87.0
arz_Arab	89.0	hau_Latn	86.0	mal_Mlym	90.0	tam_Taml	70.0
asm_Beng	87.0	heb_Hebr	85.0	mar_Deva	90.0	taq_Latn	51.0
ast_Latn	85.0	hin_Deva	91.0	min_Latn	83.0	tat_Cyrl	90.0
awa_Deva	78.0	hne_Deva	88.0	mkd_Cyrl	91.0	tel_Telu	88.0
ayr_Latn	51.0	hrv_Latn	91.0	mlt_Latn	88.0	tgk_Cyrl	90.0
azb_Arab	83.0	hun_Latn	92.0	mni_Beng	57.0	tgl_Latn	1.0
azj_Latn	90.0	hye_Armn	74.0	mos_Latn	49.0	tha_Thai	90.0
bak_Cyrl	89.0	ibo_Latn	86.0	mri_Latn	85.0	tir_Ethi	80.0
bam_Latn	59.0	ilo_Latn	65.0	mya_Mymr	86.0	tpi_Latn	81.0
ban_Latn	81.0	ind_Latn	87.0	nld_Latn	92.0	tsn_Latn	83.0
bel_Cyrl	91.0	isl_Latn	89.0	nno_Latn	91.0	tso_Latn	68.0
bem_Latn	60.0	ita_Latn	91.0	nob_Latn	94.0	tuk_Latn	77.0
ben_Beng	89.0	jav_Latn	81.0	npi_Deva	83.0	tum_Latn	80.0
bho_Deva	89.0	jpn_Jpan	90.0	nso_Latn	86.0	tur_Latn	91.0
bjn_Arab	57.0	kab_Latn	50.0	nus_Latn	44.0	twi_Latn	74.0
bjn_Latn	84.0	kac_Latn	56.0	nya_Latn	88.0	uig_Arab	77.0
bos_Latn	90.0	kam_Latn	51.0	oci_Latn	89.0	ukr_Cyrl	92.0
bug_Latn	57.0	kan_Knda	89.0	pag_Latn	58.0	umb_Latn	52.0
bul_Cyrl	85.0	kas_Arab	67.0	pan_Guru	91.0	urd_Arab	89.0
cat_Latn	93.0	kas_Deva	73.0	pap_Latn	87.0	uzn_Latn	90.0
ceb_Latn	36.0	kat_Geor	64.0	pbt_Arab	89.0	vec_Latn	89.0
ces_Latn	92.0	kaz_Cyrl	91.0	pes_Arab	89.0	vie_Latn	91.0
cjk_Latn	53.0	kbp_Latn	38.0	plt_Latn	41.0	war_Latn	65.0
ckb_Arab	87.0	kea_Latn	72.0	pol_Latn	93.0	wol_Latn	49.0
crh_Latn	86.0	khk_Cyrl	86.0	por_Latn	93.0	xho_Latn	86.0
cym_Latn	89.0	khm_Khmr	87.0	prs_Arab	91.0	ydd_Hebr	83.0
dan_Latn	93.0	kik_Latn	50.0	quy_Latn	62.0	yor_Latn	74.0
deu_Latn	93.0	kin_Latn	84.0	ron_Latn	93.0	yue_Hant	86.0
dik_Latn	46.0	kir_Cyrl	90.0	run_Latn	80.0	zho_Hans	90.0
dyu_Latn	56.0	kmb_Latn	59.0	rus_Cyrl	82.0	zho_Hant	81.0
ell_Grek	93.0	kmr_Latn	85.0	sag_Latn	74.0	zsm_Latn	90.0
eng_Latn	92.0	knc_Arab	80.0	san_Deva	75.0	zul_Latn	89.0
epo_Latn	93.0	knc_Latn	47.0	scn_Latn	87.0		
est_Latn	93.0	kon_Latn	62.0	shn_Mymr	70.0		
eus_Latn	75.0	kor_Hang	91.0	sin_Sinh	88.0		
ewe_Latn	65.0	lao_Laoo	88.0	slk_Latn	93.0		

Figure 9: The accuracy score of AYA on XStoryCloze task across 198 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	51.4	fao_Latn	76.0	lij_Latn	72.0	slv_Latn	86.2
ace_Latn	69.4	fij_Latn	56.8	lim_Latn	82.6	smo_Latn	76.8
acm_Arab	79.2	fin_Latn	85.0	lin_Latn	54.4	sna_Latn	72.8
acq_Arab	80.0	fon_Latn	50.6	lit_Latn	85.0	snd_Arab	77.2
aeb_Arab	70.8	fra_Latn	88.8	lmo_Latn	79.8	som_Latn	75.2
afr_Latn	86.4	fur_Latn	72.8	ltg_Latn	62.4	sot_Latn	79.4
ajp_Arab	78.4	fuv_Latn	51.0	ltz_Latn	81.2	spa_Latn	88.4
aka_Latn	58.6	gaz_Latn	56.8	lua_Latn	54.8	srd_Latn	78.2
als_Latn	84.0	gla_Latn	71.6	lug_Latn	61.0	srp_Cyrl	85.4
amh_Ethi	76.2	gle_Latn	74.0	luo_Latn	50.6	ssw_Latn	69.0
apc_Arab	75.6	glg_Latn	87.4	lus_Latn	58.4	sun_Latn	84.4
arb_Arab	82.0	grn_Latn	55.0	lvs_Latn	79.2	swe_Latn	87.0
ars_Arab	80.2	guj_Gujr	84.2	mag_Deva	78.2	swh_Latn	82.2
ary_Arab	75.6	hat_Latn	83.0	mai_Deva	79.4	szl_Latn	76.8
arz_Arab	80.4	hau_Latn	78.2	mal_Mlym	82.4	tam_Taml	82.8
asm_Beng	80.2	heb_Hebr	84.4	mar_Deva	84.6	taq_Latn	49.6
ast_Latn	80.0	hin_Deva	84.4	min_Latn	72.8	tat_Cyrl	81.8
awa_Deva	68.6	hne_Deva	81.6	mkd_Cyrl	85.4	tel_Telu	81.8
ayr_Latn	51.4	hrv_Latn	80.4	mlt_Latn	74.0	tgk_Cyrl	81.2
azb_Arab	70.8	hun_Latn	85.4	mni_Beng	53.2	tgl_Latn	79.4
azj_Latn	84.4	hye_Armn	83.8	mos_Latn	50.8	tha_Thai	81.6
bak_Cyrl	79.0	ibo_Latn	75.0	mri_Latn	71.2	tir_Ethi	70.2
bam_Latn	51.6	ilo_Latn	62.6	mya_Mymr	80.4	tpi_Latn	64.6
ban_Latn	67.6	ind_Latn	86.8	nld_Latn	86.8	tsn_Latn	73.2
bel_Cyrl	85.2	isl_Latn	80.2	nno_Latn	84.0	tso_Latn	56.8
bem_Latn	55.6	ita_Latn	88.0	nob_Latn	86.0	tuk_Latn	66.6
ben_Beng	84.2	jav_Latn	80.8	npi_Deva	79.2	tum_Latn	66.8
bho_Deva	81.4	jpn_Jpan	83.4	nso_Latn	76.6	tur_Latn	85.4
bjn_Arab	50.8	kab_Latn	50.4	nus_Latn	50.8	twi_Latn	60.2
bjn_Latn	77.4	kac_Latn	51.4	nya_Latn	78.2	uig_Arab	65.6
bos_Latn	82.2	kam_Latn	50.8	oci_Latn	81.6	ukr_Cyrl	86.6
bug_Latn	55.4	kan_Knda	82.6	pag_Latn	57.0	umb_Latn	51.4
bul_Cyrl	87.4	kas_Arab	56.2	pan_Guru	86.4	urd_Arab	86.0
cat_Latn	86.8	kas_Deva	65.0	pap_Latn	72.8	uzn_Latn	83.0
ceb_Latn	78.8	kat_Geor	61.0	pbt_Arab	82.0	vec_Latn	83.6
ces_Latn	85.8	kaz_Cyrl	82.0	pes_Arab	80.2	vie_Latn	84.4
cjk_Latn	52.8	kbp_Latn	51.0	plt_Latn	77.4	war_Latn	71.4
ckb_Arab	73.4	kea_Latn	62.2	pol_Latn	85.4	wol_Latn	50.2
crh_Latn	78.8	khk_Cyrl	71.8	por_Latn	88.4	xho_Latn	75.6
cym_Latn	77.4	khm_Khmr	78.2	prs_Arab	82.8	ydd_Hebr	75.6
dan_Latn	85.4	kik_Latn	50.4	quy_Latn	56.4	yor_Latn	61.8
deu_Latn	89.8	kin_Latn	69.8	ron_Latn	86.0	yue_Hant	78.2
dik_Latn	50.0	kir_Cyrl	82.8	run_Latn	68.4	zho_Hans	83.6
dyu_Latn	52.6	kmb_Latn	53.2	rus_Cyrl	85.0	zho_Hant	78.0
ell_Grek	85.0	kmr_Latn	74.0	sag_Latn	61.0	zsm_Latn	84.4
eng_Latn	89.6	knc_Arab	70.4	san_Deva	71.0	zul_Latn	75.6
epo_Latn	88.0	knc_Latn	50.4	scn_Latn	82.4		
est_Latn	84.0	kon_Latn	52.6	shn_Mymr	63.0		
eus_Latn	77.6	kor_Hang	82.4	sin_Sinh	78.8		
ewe_Latn	52.0	lao_Laoo	81.4	slk_Latn	84.8		

Figure 10: The accuracy score of AYA on XCOPA task across 198 languages.

Language	Performance	Language	performance	Language	performance	Language	performance
ace_Arab	64.37	fao_Latn	71.9	lij_Latn	71.66	slv_Latn	79.78
ace_Latn	77.41	fij_Latn	69.66	lim_Latn	73.69	smo_Latn	74.51
acm_Arab	75.45	fin_Latn	78.02	lin_Latn	68.12	sna_Latn	77.94
acq_Arab	77.03	fon_Latn	51.2	lit_Latn	78.46	snd_Arab	76.41
aeb_Arab	73.01	fra_Latn	79.58	lmo_Latn	73.33	som_Latn	75.27
afr_Latn	76.13	fur_Latn	73.95	ltg_Latn	69.92	sot_Latn	73.97
ajp_Arab	77.49	fuv_Latn	57.01	ltz_Latn	77.84	spa_Latn	80.0
aka_Latn	70.8	gaz_Latn	68.7	lua_Latn	66.41	srd_Latn	78.52
als_Latn	76.37	gla_Latn	76.45	lug_Latn	65.67	srp_Cyrl	76.99
amh_Ethi	69.16	gle_Latn	73.63	luo_Latn	58.8	ssw_Latn	69.32
apc_Arab	76.77	glg_Latn	77.01	lus_Latn	68.38	sun_Latn	81.28
arb_Arab	78.76	grn_Latn	55.41	lvs_Latn	77.6	swe_Latn	74.21
ars_Arab	76.83	guj_Gujr	73.77	mag_Deva	69.5	swh_Latn	75.41
ary_Arab	75.33	hat_Latn	75.83	mai_Deva	70.8	szl_Latn	71.04
arz_Arab	77.88	hau_Latn	79.44	mal_Mlym	70.12	tam_Taml	77.45
asm_Beng	74.97	heb_Hebr	75.19	mar_Deva	76.01	taq_Latn	62.0
ast_Latn	72.83	hin_Deva	74.79	min_Latn	74.93	tat_Cyrl	76.59
awa_Deva	68.46	hne_Deva	70.76	mkd_Cyrl	75.19	tel_Telu	77.74
ayr_Latn	53.23	hrv_Latn	75.29	mlt_Latn	71.64	tgk_Cyrl	76.63
azb_Arab	73.03	hun_Latn	76.83	mni_Beng	53.55	tgl_Latn	77.86
azj_Latn	80.34	hye_Armn	77.33	mos_Latn	51.52	tha_Thai	79.08
bak_Cyrl	74.97	ibo_Latn	75.15	mri_Latn	80.16	tir_Ethi	69.0
bam_Latn	65.55	ilo_Latn	68.52	mya_Mymr	73.81	tpi_Latn	73.93
ban_Latn	74.71	ind_Latn	78.46	nld_Latn	76.63	tsn_Latn	74.45
bel_Cyrl	78.66	isl_Latn	72.06	nno_Latn	77.56	tso_Latn	64.03
bem_Latn	62.42	ita_Latn	79.56	nob_Latn	75.13	tuk_Latn	67.45
ben_Beng	79.2	jav_Latn	78.08	npi_Deva	73.95	tum_Latn	67.74
bho_Deva	70.64	jpn_Jpan	77.66	nso_Latn	76.21	tur_Latn	79.3
bjn_Arab	64.33	kab_Latn	53.77	nus_Latn	50.92	twi_Latn	71.38
bjn_Latn	78.28	kac_Latn	67.78	nya_Latn	72.93	uig_Arab	60.78
bos_Latn	75.83	kam_Latn	52.38	oci_Latn	76.73	ukr_Cyrl	80.32
bug_Latn	59.42	kan_Knda	75.51	pag_Latn	63.61	umb_Latn	55.87
bul_Cyrl	79.42	kas_Arab	61.98	pan_Guru	71.96	urd_Arab	75.35
cat_Latn	80.04	kas_Deva	59.2	pap_Latn	74.35	uzn_Latn	78.14
ceb_Latn	76.55	kat_Geor	71.54	pbt_Arab	76.65	vec_Latn	77.03
ces_Latn	78.76	kaz_Cyrl	76.65	pes_Arab	80.98	vie_Latn	77.09
cjk_Latn	58.5	kbp_Latn	47.43	plt_Latn	77.98	war_Latn	74.45
ckb_Arab	73.79	kea_Latn	61.84	pol_Latn	78.28	wol_Latn	56.95
crh_Latn	71.24	khk_Cyrl	76.05	por_Latn	78.32	xho_Latn	69.42
cym_Latn	73.57	khm_Khmr	74.59	prs_Arab	81.06	ydd_Hebr	72.32
dan_Latn	77.07	kik_Latn	50.08	quy_Latn	54.55	yor_Latn	64.73
deu_Latn	78.42	kin_Latn	74.01	ron_Latn	73.77	yue_Hant	77.62
dik_Latn	53.43	kir_Cyrl	75.69	run_Latn	74.03	zho_Hans	76.67
dyu_Latn	63.41	kmb_Latn	57.31	rus_Cyrl	79.12	zho_Hant	72.93
ell_Grek	77.96	kmr_Latn	77.76	sag_Latn	70.78	zsm_Latn	77.29
eng_Latn	76.41	knc_Arab	69.26	san_Deva	70.44	zul_Latn	74.71
epo_Latn	76.39	knc_Latn	59.92	scn_Latn	79.42		
est_Latn	81.0	kon_Latn	62.44	shn_Mymr	67.35		
eus_Latn	76.11	kor_Hang	77.09	sin_Sinh	76.25		
ewe_Latn	58.66	lao_Laoo	73.67	slk_Latn	77.35		

Figure 11: The accuracy score of AYA on B-NLI task across 198 languages.

H Full Results for XLM-R base

Language	Performance	Language	Performance	Language	Performance	Language	Performance
ace_Arab	37.29	fao_Latn	52.85	lij_Latn	56.95	slv_Latn	75.05
ace_Latn	49.72	fij_Latn	37.37	lim_Latn	58.78	smo_Latn	37.35
acm_Arab	66.37	fin_Latn	75.65	lin_Latn	37.84	sna_Latn	38.74
acq_Arab	68.5	fon_Latn	37.5	lit_Latn	75.31	snd_Arab	71.34
aeb_Arab	56.19	fra_Latn	78.54	lmo_Latn	57.58	som_Latn	61.86
afr_Latn	77.35	fur_Latn	56.33	ltg_Latn	53.75	sot_Latn	38.36
ajp_Arab	65.61	fuv_Latn	39.34	ltz_Latn	46.31	spa_Latn	79.92
aka_Latn	38.78	gaz_Latn	45.81	lua_Latn	37.07	srd_Latn	58.2
als_Latn	76.61	gla_Latn	62.53	lug_Latn	39.82	srp_Cyrl	77.96
amh_Ethi	64.83	gle_Latn	67.45	luo_Latn	37.66	ssw_Latn	39.3
apc_Arab	61.9	glg_Latn	79.5	lus_Latn	40.1	sun_Latn	64.45
arb_Arab	73.81	grn_Latn	40.62	lvs_Latn	74.11	swe_Latn	79.08
ars_Arab	68.6	guj_Gujr	71.96	mag_Deva	63.11	swh_Latn	67.96
ary_Arab	59.76	hat_Latn	41.56	mai_Deva	57.94	szl_Latn	59.16
arz_Arab	67.66	hau_Latn	61.72	mal_Mlym	72.97	tam_Taml	72.61
asm_Beng	64.89	heb_Hebr	76.55	mar_Deva	72.83	taq_Latn	38.46
ast_Latn	62.1	hin_Deva	75.57	min_Latn	51.36	tat_Cyrl	43.59
awa_Deva	58.0	hne_Deva	59.96	mkd_Cyrl	77.19	tel_Telu	71.82
ayr_Latn	38.8	hrv_Latn	77.01	mlt_Latn	39.66	tgk_Cyrl	37.76
azb_Arab	42.02	hun_Latn	76.75	mni_Beng	35.87	tgl_Latn	71.96
azj_Latn	74.09	hye_Armn	74.23	mos_Latn	37.01	tha_Thai	73.67
bak_Cyrl	43.55	ibo_Latn	38.22	mri_Latn	38.0	tir_Ethi	42.5
bam_Latn	41.24	ilo_Latn	39.24	mya_Mymr	66.05	tpi_Latn	44.81
ban_Latn	44.39	ind_Latn	79.22	nld_Latn	77.62	tsn_Latn	39.22
bel_Cyrl	75.45	isl_Latn	70.78	nno_Latn	70.86	tso_Latn	39.44
bem_Latn	38.06	ita_Latn	78.7	nob_Latn	80.38	tuk_Latn	47.78
ben_Beng	71.84	jav_Latn	67.6	npi_Deva	71.74	tum_Latn	37.21
bho_Deva	66.15	jpn_Jpan	69.46	nso_Latn	38.58	tur_Latn	75.43
bjn_Arab	36.83	kab_Latn	35.73	nus_Latn	38.12	twi_Latn	38.24
bjn_Latn	58.18	kac_Latn	38.26	nya_Latn	39.4	uig_Arab	60.24
bos_Latn	77.35	kam_Latn	34.35	oci_Latn	64.73	ukr_Cyrl	77.74
bug_Latn	41.9	kan_Knda	72.95	pag_Latn	39.66	umb_Latn	34.69
bul_Cyrl	79.02	kas_Arab	44.43	pan_Guru	71.92	urd_Arab	73.63
cat_Latn	78.34	kas_Deva	45.61	pap_Latn	58.26	uzn_Latn	70.42
ceb_Latn	52.18	kat_Geor	54.31	pbt_Arab	71.32	vec_Latn	70.06
ces_Latn	78.34	kaz_Cyrl	70.52	pes_Arab	74.57	vie_Latn	77.11
cjk_Latn	39.02	kbp_Latn	37.11	plt_Latn	61.18	war_Latn	43.19
ckb_Arab	41.48	kea_Latn	48.32	pol_Latn	76.67	wol_Latn	39.44
crh_Latn	58.58	khk_Cyrl	70.28	por_Latn	80.02	xho_Latn	46.83
cym_Latn	71.66	khm_Khmr	69.4	prs_Arab	76.47	ydd_Hebr	67.9
dan_Latn	79.36	kik_Latn	38.54	quy_Latn	37.94	yor_Latn	36.67
deu_Latn	77.39	kin_Latn	37.8	ron_Latn	78.64	yue_Hant	66.21
dik_Latn	39.8	kir_Cyrl	68.86	run_Latn	38.2	zho_Hans	69.64
dyu_Latn	36.63	kmb_Latn	36.15	rus_Cyrl	78.74	zho_Hant	58.74
ell_Grek	77.58	kmr_Latn	64.67	sag_Latn	38.46	zsm_Latn	78.08
eng_Latn	84.09	knc_Arab	45.87	san_Deva	62.79	zul_Latn	45.89
epo_Latn	75.07	knc_Latn	38.62	scn_Latn	54.29		
est_Latn	74.65	kon_Latn	37.49	shn_Mymr	35.15		
eus_Latn	63.75	kor_Hang	73.47	sin_Sinh	71.78		
ewe_Latn	38.66	lao_Laoo	73.15	slk_Latn	78.16		

Figure 12: The accuracy score of XLM-R base on XNLI task across 196 languages.

Language	Performance	Language	Performance	Language	Performance	Language	Performance
ace_Arab	51.62	fao_Latn	58.7	lij_Latn	57.97	som_Latn	56.19
ace_Latn	56.45	fij_Latn	55.06	lim_Latn	62.21	sot_Latn	53.34
acm_Arab	61.68	fin_Latn	73.66	lin_Latn	54.4	spa_Latn	74.39
acq_Arab	67.44	fon_Latn	51.09	lit_Latn	70.62	srd_Latn	61.02
aeb_Arab	58.5	fra_Latn	72.27	lmo_Latn	60.42	srp_Cyrl	70.81
afr_Latn	72.14	fur_Latn	58.44	ltg_Latn	54.73	ssw_Latn	54.6
ajp_Arab	67.57	fuv_Latn	52.35	ltz_Latn	53.94	sun_Latn	62.14
aka_Latn	53.81	gaz_Latn	56.59	lua_Latn	49.9	swe_Latn	75.18
als_Latn	68.03	gla_Latn	59.83	lug_Latn	52.02	swh_Latn	63.73
amh_Ethi	60.56	gle_Latn	62.01	luo_Latn	56.39	szl_Latn	62.48
apc_Arab	65.12	glg_Latn	73.26	lus_Latn	54.73	tam_Taml	63.53
arb_Arab	70.62	grn_Latn	56.06	lvs_Latn	68.03	taq_Latn	51.42
ars_Arab	69.09	guj_Gujr	64.26	mag_Deva	62.67	tat_Cyrl	53.14
ary_Arab	63.34	hat_Latn	55.06	mai_Deva	59.03	tel_Telu	64.99
arz_Arab	65.06	hau_Latn	57.11	mal_Mlym	67.5	tgk_Cyrl	55.06
asm_Beng	58.24	heb_Hebr	71.01	min_Latn	57.18	tgl_Latn	64.13
ast_Latn	66.38	hin_Deva	69.16	mkd_Cyrl	71.94	tha_Thai	75.31
awa_Deva	58.9	hne_Deva	60.75	mlt_Latn	51.42	tir_Ethi	51.03
ayr_Latn	53.14	hrv_Latn	73.06	mni_Beng	50.23	tpi_Latn	57.97
azb_Arab	52.08	hun_Latn	73.26	mos_Latn	52.95	tsn_Latn	50.83
azj_Latn	69.82	hye_Armn	68.3	mri_Latn	50.96	tso_Latn	51.62
bak_Cyrl	55.72	ibo_Latn	49.97	nld_Latn	74.45	tuk_Latn	53.94
bam_Latn	52.95	ilo_Latn	54.8	nno_Latn	70.48	tum_Latn	52.42
ban_Latn	55.46	ind_Latn	76.77	nob_Latn	76.04	tur_Latn	73.46
bel_Cyrl	69.16	isl_Latn	66.84	nso_Latn	53.94	twi_Latn	52.88
bem_Latn	51.75	ita_Latn	73.79	nus_Latn	53.01	uig_Arab	60.23
ben_Beng	64.0	jav_Latn	62.41	nya_Latn	53.54	ukr_Cyrl	74.06
bho_Deva	63.86	jpn_Jpan	74.78	oci_Latn	63.67	umb_Latn	52.08
bjn_Arab	51.42	kab_Latn	52.02	pag_Latn	52.95	urd_Arab	66.78
bjn_Latn	59.1	kac_Latn	53.28	pap_Latn	59.23	uzn_Latn	63.27
bos_Latn	69.36	kam_Latn	50.83	pbt_Arab	64.39	vec_Latn	63.93
bug_Latn	53.21	kan_Knda	64.92	pes_Arab	69.76	vie_Latn	73.99
bul_Cyrl	74.59	kas_Arab	52.88	plt_Latn	55.72	war_Latn	54.93
cat_Latn	72.27	kas_Deva	58.44	pol_Latn	72.53	wol_Latn	51.49
ceb_Latn	57.58	kat_Geor	56.25	por_Latn	73.92	xho_Latn	54.07
ces_Latn	73.06	kaz_Cyrl	68.63	prs_Arab	70.88	ydd_Hebr	58.24
cjk_Latn	51.75	kbp_Latn	49.37	quy_Latn	50.36	yor_Latn	51.09
ckb_Arab	51.42	kea_Latn	56.98	ron_Latn	72.14	yue_Hant	71.21
crh_Latn	59.43	khk_Cyrl	69.69	run_Latn	47.85	zho_Hans	76.9
cym_Latn	61.35	khm_Khmr	65.06	rus_Cyrl	74.72	zho_Hant	72.47
dan_Latn	75.78	kik_Latn	53.28	sag_Latn	50.56	zsm_Latn	75.45
deu_Latn	72.87	kin_Latn	52.95	san_Deva	63.34	zul_Latn	55.53
dik_Latn	54.2	kir_Cyrl	65.98	scn_Latn	60.82		
dyu_Latn	52.68	kmb_Latn	52.35	shn_Mymr	46.33		
ell_Grek	67.44	kmr_Latn	61.48	sin_Sinh	67.04		
eng_Latn	80.34	knc_Arab	55.99	slk_Latn	73.4		
epo_Latn	69.76	knc_Latn	52.88	slv_Latn	71.34		
est_Latn	71.61	kon_Latn	53.94	smo_Latn	51.69		
eus_Latn	67.44	kor_Hang	69.89	sna_Latn	54.33		
ewe_Latn	51.95	lao_Laoo	64.53	snd_Arab	63.0		

Figure 13: The accuracy score of XLM-R base on XStoryCloze task across 196 languages.

Language	Performance	Language	Performance	Language	Performance	Language	Performance
ace_Arab	67.3	fao_Latn	84.8	lij_Latn	82.95	slv_Latn	88.3
ace_Latn	82.6	fij_Latn	71.5	lim_Latn	85.5	smo_Latn	73.7
acm_Arab	85.9	fin_Latn	84.15	lin_Latn	74.8	sna_Latn	75.25
acq_Arab	85.9	fon_Latn	66.35	lit_Latn	87.45	snd_Arab	84.3
aeb_Arab	84.2	fra_Latn	89.9	lmo_Latn	84.9	som_Latn	83.0
afr_Latn	91.05	fur_Latn	83.6	ltg_Latn	80.8	sot_Latn	76.3
ajp_Arab	85.1	fuv_Latn	70.8	ltz_Latn	82.3	spa_Latn	89.4
aka_Latn	70.2	gaz_Latn	74.0	lua_Latn	71.55	srd_Latn	85.2
als_Latn	90.05	gla_Latn	84.6	lug_Latn	72.6	srp_Cyrl	89.65
amh_Ethi	81.5	gle_Latn	85.1	luo_Latn	71.3	ssw_Latn	74.15
apc_Arab	84.85	glg_Latn	89.8	lus_Latn	75.1	sun_Latn	87.35
arb_Arab	86.25	grn_Latn	75.0	lvs_Latn	86.2	swe_Latn	90.35
ars_Arab	85.25	guj_Gujr	83.95	mag_Deva	84.35	swh_Latn	86.95
ary_Arab	84.35	hat_Latn	84.4	mai_Deva	80.65	szl_Latn	85.25
arz_Arab	84.85	hau_Latn	83.8	mal_Mlym	78.75	tam_Taml	81.75
asm_Beng	80.35	heb_Hebr	86.95	mar_Deva	82.1	taq_Latn	70.2
ast_Latn	85.85	hin_Deva	86.1	min_Latn	85.5	tat_Cyrl	77.65
awa_Deva	78.9	hne_Deva	82.6	mkd_Cyrl	89.85	tel_Telu	83.65
ayr_Latn	68.95	hrv_Latn	89.5	mlt_Latn	79.9	tgk_Cyrl	75.5
azb_Arab	69.35	hun_Latn	87.3	mni_Beng	71.0	tgl_Latn	89.8
azj_Latn	83.8	hye_Armn	86.05	mos_Latn	66.6	tha_Thai	86.25
bak_Cyrl	75.8	ibo_Latn	72.55	mri_Latn	75.05	tir_Ethi	76.15
bam_Latn	72.8	ilo_Latn	78.9	mya_Mymr	78.85	tpi_Latn	73.7
ban_Latn	81.0	ind_Latn	91.0	nld_Latn	89.55	tsn_Latn	76.35
bel_Cyrl	87.55	isl_Latn	86.6	nno_Latn	84.2	tso_Latn	72.55
bem_Latn	70.75	ita_Latn	90.45	nob_Latn	91.05	tuk_Latn	71.65
ben_Beng	84.85	jav_Latn	87.6	npi_Deva	84.9	tum_Latn	70.75
bho_Deva	80.55	jpn_Jpan	81.35	nso_Latn	76.85	tur_Latn	83.35
bjn_Arab	66.65	kab_Latn	72.4	nus_Latn	67.3	twi_Latn	70.9
bjn_Latn	86.6	kac_Latn	71.25	nya_Latn	75.85	uig_Arab	77.95
bos_Latn	89.5	kam_Latn	62.85	oci_Latn	88.4	ukr_Cyrl	89.3
bug_Latn	78.2	kan_Knda	83.85	pag_Latn	78.45	umb_Latn	63.6
bul_Cyrl	89.95	kas_Arab	71.1	pan_Guru	84.95	urd_Arab	85.15
cat_Latn	90.65	kas_Deva	70.55	pap_Latn	84.95	uzn_Latn	82.35
ceb_Latn	84.85	kat_Geor	78.25	pbt_Arab	83.2	vec_Latn	88.05
ces_Latn	89.7	kaz_Cyrl	82.7	pes_Arab	85.55	vie_Latn	89.85
cjk_Latn	69.35	kbp_Latn	65.65	plt_Latn	82.4	war_Latn	80.7
ckb_Arab	72.35	kea_Latn	82.7	pol_Latn	88.2	wol_Latn	71.95
crh_Latn	81.8	khk_Cyrl	80.4	por_Latn	89.8	xho_Latn	81.1
cym_Latn	88.0	khm_Khmr	86.45	prs_Arab	86.9	ydd_Hebr	85.3
dan_Latn	90.0	kik_Latn	69.45	quy_Latn	67.55	yor_Latn	69.45
deu_Latn	89.55	kin_Latn	71.65	ron_Latn	89.95	yue_Hant	80.0
dik_Latn	69.3	kir_Cyrl	79.25	run_Latn	72.0	zho_Hans	79.65
dyu_Latn	66.7	kmb_Latn	65.0	rus_Cyrl	89.3	zho_Hant	69.5
ell_Grek	88.8	kmr_Latn	84.2	sag_Latn	70.65	zsm_Latn	91.25
eng_Latn	93.15	knc_Arab	72.8	san_Deva	73.65	zul_Latn	82.3
epo_Latn	89.5	knc_Latn	66.9	scn_Latn	84.8		
est_Latn	86.15	kon_Latn	71.25	shn_Mymr	66.6		
eus_Latn	78.35	kor_Hang	82.15	sin_Sinh	82.95		
ewe_Latn	71.45	lao_Laoo	86.6	slk_Latn	89.9		

Figure 14: The accuracy score of XLM-R base on PAWS-X task across 196 languages.

I ChrF++ scores per language

Language	Chrf++ score	Language	Chrf++ score	Language	Chrf++ score	Language	Chrf++ score
ace_Arab	18.38	fij_Latn	45.96	lim_Latn	44.76	smo_Latn	50.43
ace_Latn	42.42	fin_Latn	51.3	lin_Latn	50.46	sna_Latn	43.66
acm_Arab	39.32	fon_Latn	20.6	lit_Latn	51.6	snd_Arab	49.86
acq_Arab	42.23	fra_Latn	67.2	lmo_Latn	34.74	som_Latn	46.02
aeb_Arab	36.3	fur_Latn	55.68	ltg_Latn	44.87	sot_Latn	44.44
afr_Latn	65.22	fuv_Latn	24.62	ltz_Latn	54.1	spa_Latn	54.7
ajp_Arab	44.92	gaz_Latn	36.33	lua_Latn	36.78	srd_Latn	54.78
aka_Latn	36.34	gla_Latn	49.15	lug_Latn	39.41	srp_Cyrl	55.24
als_Latn	56.05	gle_Latn	55.47	luo_Latn	41.13	ssw_Latn	43.68
amh_Ethi	30.23	glg_Latn	58.52	lus_Latn	40.26	sun_Latn	48.42
apc_Arab	43.91	grn_Latn	34.0	lvs_Latn	46.98	swe_Latn	65.14
arb_Arab	52.77	guj_Gujr	48.15	mag_Deva	59.0	swh_Latn	62.33
ars_Arab	47.24	hat_Latn	52.06	mai_Deva	44.7	szl_Latn	49.19
ary_Arab	36.35	hau_Latn	53.77	mal_Mlym	43.29	tam_Taml	52.25
arz_Arab	43.39	heb_Hebr	52.48	mar_Deva	44.48	taq_Latn	27.69
asm_Beng	35.68	hin_Deva	56.72	min_Latn	55.33	tat_Cyrl	45.93
ast_Latn	53.02	hne_Deva	51.64	mkd_Cyrl	56.51	tel_Telu	51.88
awa_Deva	39.64	hrv_Latn	54.07	mlt_Latn	63.73	tgk_Cyrl	46.24
ayr_Latn	31.75	hun_Latn	54.43	mni_Beng	38.16	tgl_Latn	59.93
azb_Arab	23.53	hye_Armn	51.02	mos_Latn	27.92	tha_Thai	39.54
azj_Latn	42.08	ibo_Latn	41.94	mri_Latn	46.87	tir_Ethi	24.48
bak_Cyrl	46.46	ilo_Latn	54.64	mya_Mymr	32.35	tpi_Latn	44.67
bam_Latn	31.33	ind_Latn	65.11	nld_Latn	57.1	tsn_Latn	47.51
ban_Latn	43.27	isl_Latn	47.98	nno_Latn	49.49	tso_Latn	54.34
bel_Cyrl	40.34	ita_Latn	56.32	nob_Latn	58.69	tuk_Latn	36.45
bem_Latn	37.32	jav_Latn	53.07	npi_Deva	45.63	tum_Latn	38.62
ben_Beng	45.48	jpn_Jpan	24.78	nso_Latn	51.5	tur_Latn	57.18
bho_Deva	38.01	kab_Latn	31.66	nus_Latn	26.86	twi_Latn	40.66
bjn_Arab	21.12	kac_Latn	38.81	nya_Latn	46.42	uig_Arab	36.37
bjn_Latn	46.01	kam_Latn	24.22	oci_Latn	59.63	ukr_Cyrl	52.2
bos_Latn	56.14	kan_Knda	49.11	pag_Latn	49.66	umb_Latn	25.47
bug_Latn	35.41	kas_Arab	32.66	pan_Guru	49.79	urd_Arab	50.53
bul_Cyrl	61.75	kas_Deva	20.7	pap_Latn	55.38	uzn_Latn	47.26
cat_Latn	64.34	kat_Geor	42.11	pbt_Arab	34.14	vec_Latn	47.31
ceb_Latn	57.88	kaz_Cyrl	44.98	pes_Arab	44.16	vie_Latn	60.12
ces_Latn	54.49	kbp_Latn	33.21	plt_Latn	51.81	war_Latn	57.14
cjk_Latn	26.38	kea_Latn	49.26	pol_Latn	49.25	wol_Latn	30.78
ckb_Arab	44.65	khk_Cyrl	38.78	por_Latn	67.53	xho_Latn	43.99
crh_Latn	41.15	khm_Khmr	35.58	prs_Arab	44.15	ydd_Hebr	35.95
cym_Latn	65.55	kik_Latn	36.27	quy_Latn	28.58	yor_Latn	29.3
dan_Latn	64.09	kin_Latn	51.13	ron_Latn	60.1	yue_Hant	16.82
deu_Latn	61.69	kir_Cyrl	41.22	run_Latn	45.04	zho_Hans	17.0
dik_Latn	22.68	kmb_Latn	26.6	rus_Cyrl	53.87	zho_Hant	11.99
dyu_Latn	17.86	kmr_Latn	40.57	sag_Latn	34.33	zsm_Latn	65.86
ell_Grek	53.36	knc_Arab	11.39	san_Deva	22.74	zul_Latn	51.69
epo_Latn	60.17	knc_Latn	23.95	scn_Latn	43.16		
est_Latn	54.95	kon_Latn	44.35	shn_Mymr	36.29		
eus_Latn	43.48	kor_Hang	33.04	sin_Sinh	40.13		
ewe_Latn	38.58	lao_Laoo	49.11	slk_Latn	55.71		
fao_Latn	49.94	lij_Latn	46.65	slv_Latn	53.92		

Figure 15: Chrf++ scores for the selected languages.

J Language coverage

	Category	Languages
XLM-R	High	arb_Arab , bul_Cyrl , dan_Latn , deu_Latn , ell_Grek , eng_Latn , fin_Latn , fra_Latn , heb_Hebr , hun_Latn , ind_Latn , ita_Latn , jpn_Jpan , kor_Hang , nld_Latn , nob_Latn , pes_Arab , pol_Latn , por_Latn , ron_Latn , rus_Cyrl , spa_Latn , tha_Thai , ukr_Cyrl , vie_Latn , zho_Hans
	Mid	als_Latn , azj_Latn , bel_Cyrl , ben_Beng , cat_Latn , ces_Latn , est_Latn , glg_Latn , hin_Deva , hrv_Latn , hye_Armn , isl_Latn , kan_Knda , kat_Geor , kaz_Cyrl , khk_Cyrl , lit_Latn , lvs_Latn , mal_Mlym , mar_Deva , mkd_Cyrl , npi_Deva , sin_Sinh , slk_Latn , slv_Latn , srp_Cyrl , swe_Latn , tam_Taml , tel_Telu , tgl_Latn , tur_Latn , urd_Arab , zho_Hant , zsm_Latn
	Low	afr_Latn , amh_Ethi , asm_Beng , bos_Latn , ckb_Arab , cym_Latn , epo_Latn , eus_Latn , gaz_Latn , gla_Latn , gle_Latn , guj_Gujr , hau_Latn , jav_Latn , khm_Khmr , kir_Cyrl , lao_Laoo , mya_Mymr , pan_Guru , pbt_Arab , plt_Latn , san_Deva , snd_Arab , som_Latn , sun_Latn , swh_Latn , uig_Arab , uzn_Latn , xho_Latn , ydd_Hebr
	Unseen	ace_Arab , ace_Latn , acm_Arab , acq_Arab , aeb_Arab , ajp_Arab , aka_Latn , apc_Arab , ars_Arab , ary_Arab , arz_Arab , ast_Latn , awa_Deva , ayr_Latn , azb_Arab , bak_Cyrl , bam_Latn , ban_Latn , bem_Latn , bho_Deva , bjn_Arab , bjn_Latn , bug_Latn , ceb_Latn , cjk_Latn , crh_Latn , dik_Latn , dyu_Latn , ewe_Latn , fao_Latn , fij_Latn , fon_Latn , fur_Latn , fuv_Latn , grn_Latn , hat_Latn , hne_Deva , ibo_Latn , ilo_Latn , kab_Latn , kac_Latn , kam_Latn , kas_Arab , kas_Deva , kbp_Latn , kea_Latn , kik_Latn , kin_Latn , kmb_Latn , kmr_Latn , knc_Arab , knc_Latn , kon_Latn , lij_Latn , lim_Latn , lin_Latn , lmo_Latn , ltg_Latn , ltz_Latn , lua_Latn , lug_Latn , luo_Latn , lus_Latn , mag_Deva , mai_Deva , min_Latn , mlt_Latn , mni_Beng , mos_Latn , mri_Latn , nno_Latn , nso_Latn , nus_Latn , nya_Latn , oci_Latn , pag_Latn , pap_Latn , prs_Arab , quy_Latn , run_Latn , sag_Latn , scn_Latn , shn_Mymr , smo_Latn , sna_Latn , sot_Latn , srd_Latn , ssw_Latn , szl_Latn , taq_Latn , tat_Cyrl , tgk_Cyrl , tir_Ethi , tpi_Latn , tsn_Latn , tso_Latn , tuk_Latn , tum_Latn , twi_Latn , umb_Latn , vec_Latn , war_Latn , wol_Latn , yor_Latn , yue_Hant , zul_Latn
BLOOMz	High	arb_Arab , cat_Latn , eng_Latn , fra_Latn , ind_Latn , por_Latn , spa_Latn , vie_Latn , zho_Hans
	Mid	ben_Beng , eus_Latn , hin_Deva , mal_Mlym , tam_Taml , urd_Arab , zho_Hant’
	Low	aka_Latn , asm_Beng , bam_Latn , bho_Deva , fon_Latn , guj_Gujr , ibo_Latn , kan_Knda , kik_Latn , kin_Latn , lin_Latn , mar_Deva , npi_Deva , nso_Latn , sot_Latn , swh_Latn , tel_Telu , wol_Latn , xho_Latn , yor_Latn , zul_Latn
	Unseen	ace_Arab , ace_Latn , acm_Arab , acq_Arab , aeb_Arab , afr_Latn , ajp_Arab , apc_Arab , ars_Arab , ary_Arab , arz_Arab , ast_Latn , awa_Deva , ayr_Latn , azb_Arab , azj_Latn , ban_Latn , bem_Latn , bjn_Arab , bjn_Latn , bos_Latn , bug_Latn , ceb_Latn , ces_Latn , cjk_Latn , ckb_Arab , crh_Latn , cym_Latn , dan_Latn , deu_Latn , dik_Latn , dyu_Latn , epo_Latn , est_Latn , ewe_Latn , fao_Latn , fij_Latn , fin_Latn , fur_Latn , fuv_Latn , gla_Latn , gle_Latn , glg_Latn , grn_Latn , hat_Latn , hau_Latn , hne_Deva , hrv_Latn , hun_Latn , ilo_Latn , isl_Latn , ita_Latn , jav_Latn , kab_Latn , kac_Latn , kam_Latn , kas_Arab , kas_Deva , knc_Arab , knc_Latn , kbp_Latn , kea_Latn , kmb_Latn , kmr_Latn , kon_Latn , lij_Latn , lim_Latn , lit_Latn , lmo_Latn , ltg_Latn , ltz_Latn , lua_Latn , lug_Latn , luo_Latn , lus_Latn , lvs_Latn , mag_Deva , mai_Deva , min_Latn , plt_Latn , mlt_Latn , mni_Beng , mos_Latn , mri_Latn , nld_Latn , nno_Latn , nob_Latn , nus_Latn , nya_Latn , oci_Latn , gaz_Latn , pag_Latn , pap_Latn , pes_Arab , pol_Latn , prs_Arab , pbt_Arab , quy_Latn , ron_Latn , run_Latn , sag_Latn , san_Deva , scn_Latn , slk_Latn , slv_Latn , smo_Latn , sna_Latn , snd_Arab , som_Latn , als_Latn , srd_Latn , ssw_Latn , sun_Latn , swe_Latn , szl_Latn , tgl_Latn , taq_Latn , tpi_Latn , tsn_Latn , tso_Latn , tuk_Latn , tum_Latn , tur_Latn , twi_Latn , uig_Arab , umb_Latn , uzn_Latn , vec_Latn , war_Latn , yue_Hant , zsm_Latn

Table 13: The languages covered during pretraining of each of the MLMs categorized by the amount of data that was seen for them during pretraining.

	Category	Languages
AYA	High	hye_Armn , kan_Knda , tur_Latn , ita_Latn , nld_Latn , pol_Latn , por_Latn , isl_Latn , fra_Latn , deu_Latn , spa_Latn , rus_Cyrl , eng_Latn
	Mid	est_Latn , ben_Beng , mar_Deva , slv_Latn , lit_Latn , heb_Hebr , zsm_Latn , cat_Latn , tha_Thai , kor_Hang , slk_Latn , hin_Deva , bul_Cyrl , nob_Latn , fin_Latn , dan_Latn , hun_Latn , ukr_Cyrl , ell_Grek , ron_Latn , swe_Latn , arb_Arab , pes_Arab , zho_Hans , ces_Latn
	Low	hat_Latn , kor_Hang , xho_Latn , ibo_Latn , lao_Laoo , mri_Latn , smo_Latn , ckb_Arab , amh_Ethi , nya_Latn , hau_Latn , plt_Latn , pbt_Arab , gla_Latn , sun_Latn , jpn_Jpan , sot_Latn , ceb_Latn , pan_Guru , gle_Latn , kir_Cyrl , epo_Latn , sin_Sinh , guj_Gujr , yor_Latn , tgk_Cyrl , snd_Arab , mya_Mymr , kaz_Cyrl , khm_Khmr , som_Latn , swh_Latn , ydd_Hebr , uzn_Latn , hun_Latn , mlt_Latn , eus_Latn , bel_Cyrl , kat_Geor , mkd_Cyrl , mal_Mlym , khk_Cyrl , tha_Thai , afr_Latn , ukr_Cyrl , ltz_Latn , tel_Telu , urd_Arab , lit_Latn , npi_Deva , srp_Cyrl , tam_Taml , cym_Latn , als_Latn , glg_Latn , azj_Latn , lvs_Latn
	Unseen	ace_Arab , ace_Latn , acm_Arab , acq_Arab , aeb_Arab , ajp_Arab , aka_Latn , apc_Arab , ars_Arab , ary_Arab , arz_Arab , asm_Beng , ast_Latn , awa_Deva , ayr_Latn , azb_Arab , bak_Cyrl , bam_Latn , ban_Latn , bem_Latn , bho_Deva , bjn_Arab , bjn_Latn , bos_Latn , bug_Latn , cjk_Latn , crh_Latn , dik_Latn , dyu_Latn , ewe_Latn , fao_Latn , fij_Latn , fon_Latn , fur_Latn , fuv_Latn , grn_Latn , hne_Deva , hrv_Latn , ilo_Latn , kab_Latn , kac_Latn , kam_Latn , kas_Arab , kas_Deva , knc_Arab , knc_Latn , kbp_Latn , kea_Latn , kik_Latn , kin_Latn , kmb_Latn , kmr_Latn , kon_Latn , lij_Latn , lim_Latn , lin_Latn , lmo_Latn , ltg_Latn , lua_Latn , lug_Latn , luo_Latn , lus_Latn , mag_Deva , mai_Deva , min_Latn , mni_Beng , mos_Latn , nno_Latn , nso_Latn , nus_Latn , oci_Latn , gaz_Latn , pag_Latn , pap_Latn , prs_Arab , quy_Latn , run_Latn , sag_Latn , san_Deva , scn_Latn , shn_Mymr , srd_Latn , ssw_Latn , szl_Latn , tat_Cyrl , tgl_Latn , tir_Ethi , taq_Latn , tpi_Latn , tsn_Latn , tso_Latn , tuk_Latn , tum_Latn , twi_Latn , tzm_Tfng , uig_Arab , umb_Latn , vec_Latn , war_Latn , wol_Latn , yue_Hant , zho_Hant , dzo_Tibt

Table 14: The languages covered during AYA's pretraining categorized by the amount of data that was seen during pretraining.

Cross-lingual Human-Preference Alignment for Neural Machine Translation with Direct Quality Optimization

Kaden Uhlig[1], **Joern Wübker**[1], **John DeNero**[1], **Raphael Reinauer**[2]*
[1]LILT, [2]Amazon
{kaden.uhlig,joern,john}@lilt.com, raphada@amazon.de

Abstract

Reinforcement Learning from Human Feedback (RLHF) and derivative techniques like Direct Preference Optimization (DPO) are task-alignment algorithms used to repurpose general, foundational models for specific tasks. We show that applying task-alignment to neural machine translation (NMT) addresses an existing task–data mismatch in NMT, leading to improvements across all languages of a multilingual model, even when task-alignment is only applied to a subset of those languages. We do so by introducing Direct Quality Optimization (DQO), a variant of DPO leveraging a pre-trained translation quality estimation model as a proxy for human preferences, and verify the improvements with both automatic metrics and through human evaluation.

1 Introduction

For many natural language generation (NLG) tasks, aligning models to human preferences has led to large performance gains (Ziegler et al., 2020). A strong motivation for this alignment step is that much of the data on which the model was originally trained – internet text – is useful for language generation in general but does not match the desired output for the task. Neural machine translation (NMT) models have not involved alignment to human preferences, in part because of the assumption that supervised training data for NMT does match the desired output of the translation task. However, we show the existence of a mismatch between the NMT task and typical training data.

Machine translation is unusual among NLG tasks in that task-relevant supervised training data – text paired with its translation – is plentiful and publicly available. One might expect that with such a large amount of task-relevant training data, there would be no need for task-alignment. However, we identify an exhaustive list of reasons why training examples in a parallel corpus diverge from the desired output in meaningful ways (see Section 2.2).

Machine translation is also unusual in that human preference data has been collected and published for a large number of systems, and translation quality estimation (QE) is an active research area that has benefited greatly from recent advances in large language models. We introduce a method for using quality estimation models, which themselves are trained from human preference data, in order to perform NMT task alignment. Our method, Direct Quality Optimization (DQO), is a batched online variant of Direct Preference Optimization (DPO) (Rafailov et al., 2023) that uses a QE model as a proxy for human preference.

We show that DQO improves translation quality in terms of BLEU, COMET22, CometKiwi22, and BLEURT, and leads to a reduction in translation errors in a human evaluation using the Multidimensional Quality Metric framework (MQM) (Lommel et al., 2014; Freitag et al., 2021).

We make three notable observations when applying DQO to a multilingual model:

1. Task alignment increases task performance and human preference while also increasing the distance between the model's output distribution and the training data distribution.
2. Improvements carry over to held-out languages and language families, which were not contained in the data used for DQO.
3. Improvements in held-out languages are not limited to general behaviors required by the translation task (e.g. avoiding omissions), but include improving language-specific linguistic features not seen in the DQO alignment data, such as correctly transliterating named entities in Latvian.

While we attribute much of the performance in held-out languages to transfer learning of general behaviors required by the translation task

*Contributions made prior to joining Amazon.

Proceedings of the Tenth Conference on Machine Translation, pages 31–51
November 8-9, 2025

(e.g. avoiding omissions), the language-specific improvements in held-out languages cannot be explained by transfer learning.

Instead, these results suggest that DQO does not only increase the likelihood of the features present in its task alignment data, but also focuses the model on human preference features that it already learned during supervised training.

2 The Task–Data Mismatch in NMT

2.1 Task: Human-Preferred Translations

Like many NLG tasks, NMT is an open-ended problem, with multiple valid outputs for any given input, each preferred more or less by humans depending on a variety of factors, including adequacy, fluency, context, tone, style, and many other subtle features.

Because of this, the task of NMT cannot be reduced to producing valid translations, nor human-like translations, but instead requires generating human-preferred translations – those judged as at least as good as all other valid translations.

2.2 Training Data Mismatch

The supervised training data used in NMT comes from a variety of sources, each with notable differences from the task distribution of human-preferred translations.

Web Data Mining. A large portion of parallel data is mined from massive collections of web documents, using automated methods to align source and target language segments – e.g. the ParaCrawl (Bañón et al., 2020) and CCMatrix (Schwenk et al., 2021b) datasets. This process may capture human translations, text written independently in both the source and target languages on the same topic, or the output of other MT models. One prominent cause of task–data mismatch in automatically aligned sentence pairs is semantic misalignment. Kreutzer et al. (2022) found semantic misalignment in 15% (ParaCrawl) and 32% (CCMatrix) of sentence pairs in a manual quality audit.

The simplest form is complete semantic misalignment, when the source and target segments are completely unrelated. This certainly contributes to any task–data mismatch, but such pairs are easy to detect with tools such as BiCleaner (Ramírez-Sánchez et al., 2020) or reference-free quality evaluation models such as CometKiwi22 (Peter et al., 2023).

Unfortunately, slight semantic misalignments of source and target are both more prevalent and much more difficult for state-of-the-art filtering systems to detect (Meng et al., 2024). These may include subtle yet significant differences in meaning, factual differences in numbers or names, additions and omissions, and the accompanying losses in translation adequacy. In addition, these segments often still contain useful information that may help the model learn (Meng et al., 2024).

Accidental Inclusion of Machine Translated Content. Web data may also include the outputs of other machine translation models, including neural, statistical and dictionary-based methods of varying quality. The impact of training on low quality machine translations is clear, however even good NMT systems' outputs differ significantly enough from natural text that classifiers can detect machine translated text with high accuracy – and even predict which machine translation system was used to translate a given text (La Morgia et al., 2023).

Recent research suggests that up to 57% of translations mined from the web are multi-way parallel, meaning parallel translations of a segment can be found in more than two languages, and demonstrates a strong correlation between multi-way parallelism and low quality translations likely to be machine translated (Thompson et al., 2024). The authors also found that multi-way parallel translations follow a distinct distribution, focused on low-quality content typically used for search engine optimization.

Translator Skill Level. Another source of task–data mismatch in human translations is the fact that translators differ in skill level (Albir, 2017). This implies that not all human translations will be equally preferred by humans.

Achieving mean human quality in translations is not the task of NMT as defined in Section 2.1. We propose that neither is *maximum* human quality. In theory it is conceivable that humans prefer machine-generated translations over even the best human-generated translations. Therefore, we do not want finite human skill to impose an upper limit on translation quality.

Translationese. Another common issue is a phenomenon known as *translationese*, the observation that human-translated text in a given language differ in distribution from text written independently in that language. Specifically, translated text shows signs of interference from the source language's grammar, word order and word choice, as well as source-language-independent effects of the translation process itself, such as simplification

Model	Lang.	FLORES+ devtest				NTREX			
		BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
Baseline	All	0.7614	0.8741	0.8387	34.19	0.7016	0.8359	0.8099	30.31
DQO	All	**0.7790**	**0.8873**	**0.8508**	**35.31**	**0.7212**	**0.8525**	**0.8255**	**31.21**
Baseline	$\mathcal{T}$	0.7231	0.8417	0.8272	34.50	0.6677	0.8040	0.7979	32.62
DQO	$\mathcal{T}$	**0.7381**	**0.8559**	**0.8401**	**35.34**	**0.6854**	**0.8209**	**0.8137**	**33.16**
Baseline	$\mathcal{T}^c$	0.7691	0.8805	0.8410	34.13	0.7084	0.8423	0.8123	29.85
DQO	$\mathcal{T}^c$	**0.7872**	**0.8935**	**0.8529**	**35.30**	**0.7284**	**0.8588**	**0.8278**	**30.82**
Baseline	$\mathcal{R} \cap \mathcal{T}^c$	0.7802	0.8820	0.8447	36.46	0.7202	0.8432	0.8154	33.01
DQO	$\mathcal{R} \cap \mathcal{T}^c$	**0.7967**	**0.8936**	**0.8557**	**37.54**	**0.7391**	**0.8593**	**0.8307**	**34.13**
Baseline	$\mathcal{R}^c$	0.7549	0.8787	0.8364	31.17	0.6934	0.8413	0.8084	25.84
DQO	$\mathcal{R}^c$	**0.7751**	**0.8934**	**0.8493**	**32.46**	**0.7147**	**0.8581**	**0.8242**	**26.61**

Table 1: **Evaluation metrics on FLORES+ devtest and NTREX** with the NVIDIA Megatron EN-X model, before and after task-alignment using DQO. Results are shown for relevant groupings of the 30 target languages: all languages, languages used in DQO ($\mathcal{T}$), languages not used in DQO ($\mathcal{T}^c$), languages not used in DQO, but related to those used in DQO ($\mathcal{R} \cap \mathcal{T}^c$), and languages neither used nor related to the languages used in DQO ($\mathcal{R}^c$).

and avoidance of unique language features (Koppel and Ordan, 2011; Laviosa, 1998; Tirkkonen-Condit, 2004). These effects are significant enough that classification models can distinguish translated and original text with high accuracy (Baroni and Bernardini, 2005; Sominsky and Wintner, 2019), as well as identifying the source language of the text (Koppel and Ordan, 2011). As humans show a consistent preference for translations closer to the distribution of original text rather than translationese (Riley et al., 2020; Freitag et al., 2022b), this creates an inherent task–data mismatch for training data translated in the source–target direction.

Source–Target Domain Mismatch. Translation pairs in the other direction, target–source, are better aligned with human preference, as the target labels are drawn from the original text distribution rather than from translationese. Unfortunately, they suffer from another subtle source of task–data mismatch found in human translations: Source–target domain mismatch (Shen et al., 2021) is the observation that speakers of different languages tend to discuss different topics. For instance, a Cherokee newspaper is likely to report on different topics than an Icelandic newspaper would, and translations of these topics would remain representative of the Cherokee domain. This effect is especially pronounced for low-resource language pairs (Shen et al., 2021).

If one were to avoid the task–data mismatch of translationese by using only target–source translation pairs, the training data may lack key information about topics found only in the source domain. Because the task is translation from the source domain into the target language, this, too, would represent an unavoidable task–data mismatch.

3 Human Preference Learning for LLMs

Supervised data showing chat-based dialog between humans and AI assistants was, prior to the wide availability of such agents in the form of LLMs, understandably rare. Even with the advent of high quality proprietary and open-source models, which one could sample to create synthetic data, there is a fundamental task–data mismatch: the task is not to imitate an existing AI assistant, but (ideally) to train a new state-of-the-art model.

LLM training instead follows a two-step process:

1. Supervised learning on massive amounts of web data.
2. Task alignment using instruction fine-tuning and human preference learning.

In step one, the actual task for which the model is optimized is predicting the next token in documents taken from the web. This, when done at scale and with a variety of data sources, provides the model with extensive world knowledge and understanding of a wide array of styles and document types.

This is then followed by instruction fine-tuning, a comparatively brief round of supervised learning on human- or AI-labeled examples of dialogues, which brings the model's output distribution into the general neighborhood of desired behavior. Finally human preference learning, using actual human rankings aligns the model with the desired task: producing human-preferred responses to questions and dialog, while remaining helpful and harmless (Bai et al., 2022).

Direct Preference Optimization (DPO) is a preference learning algorithm that trains on preference pairs of the form (x, y_w, y_l), with x being a model input, and y_w and y_l being two potential model outputs for the input x, marked as chosen (winning) or rejected (losing) by humans during data collection (Rafailov et al., 2023), using the loss function:

$$\mathcal{L}_{\text{DPO}}(x, y_w, y_l) = \log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right), \quad (1)$$

where σ is the logistic function.

4 Direct Quality Optimization for NMT

Because of its stability and ease of use, we select DPO as the basis for our experiments with human preference learning as a form of task alignment for NMT. As a proxy for human preferences, we use the CometKiwi22 quality estimation model to score and compare multiple translations of a given source (Rei et al., 2022b). CometKiwi22 is highly multilingual and has been shown to correlate well with human preference (Kocmi et al., 2024). To verify that our method is not dependent on the specific choice of quality estimation model we conducted a brief experiment using MetricX (Juraska et al., 2023) instead of CometKiwi22 and obtained very similar results.

Our main experiments are run with the NVIDIA Megatron English–Many model[1], a 500M parameter encoder-decoder model, which supports translating from English into 30 languages[2] from 14 language families, listed in Table 2. We denote the complete list of supported target languages as $\mathcal{M}$.

Language Family	Languages (ISO 639-1)
Baltic	lt, lv
Germanic	da, **de**, nl, no, sv
Romance	**es**, fr, it, pt, ro
Slavic	bg, cs, hr, pl, **ru**, sl, uk
Uralic	et, fi, hu
Other	el, **hi**, id, ja, ko, tr, vi, **zh**

Table 2: **Target languages supported by the NVIDIA Megatron En-X model**. The category "Other" contains all languages that are the only supported representative of their language family. The languages on which we apply task alignment are denoted in boldface.

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/megatronnmt_en_any_500m

[2] The model was originally trained to support 32 languages, but we found that translating into Arabic and Slovak resulted in degenerate output.

The model's multilingual nature allows us to apply task alignment to a subset of language pairs and observe the effects on unrelated languages, with minimal risk of exposing the model to any new information in those languages.

Any improvements in those languages must either apply to all languages (such as avoiding omissions or additions), or are language specific, and can only have come from previously unused latent knowledge acquired during supervised training.

In our experiments, we selected Chinese, German, Hindi, Russian and Spanish as the target languages used during task alignment, termed $\mathcal{T} = \{de, es, hi, ru, es\}$. Let $\mathcal{T}^C = \mathcal{M} \setminus \mathcal{T}$ be the set containing the 25 target languages not represented during task alignment, R be the set of languages related to at least one language in $\mathcal{T}$ (defined as belonging to the same language family), and $\mathcal{R}^C = \mathcal{M} \setminus \mathcal{R}$ be the languages unrelated to any of the languages used in task alignment. An overview of how many languages belong to each set is shown in Table 3.

Subset	Definition	Size
$\mathcal{T}$	Languages seen in DQO	5
$\mathcal{T}^C$	Languages not seen in DQO	25
$\mathcal{R}$	Languages related to $\mathcal{T}$	19
$\mathcal{R}^C$	Languages unrelated to $\mathcal{T}$	11

Table 3: **Target languages supported by the NVIDIA Megatron EN-X model**, categorized by their relationship with the languages selected for task alignment.

As the seed dataset from which to draw source sentences for human preference learning, we use the source side of a mixture of publicly available English–German MT datasets (see Appendix A.4).

From this dataset, we sample 8000 source segments. For each source segment, we sample a target language from $\mathcal{T}$, the languages used for task alignment, and use the current policy model to sample 64 translations into that language using combined Top-K and Top-P sampling, with $K = 40$, $P = 0.8$ (Fan et al., 2018; Holtzman et al., 2020). We also add the greedy translation for each source segment, obtaining a total of 520 000 translations.

Letting the output of the CometKiwi22 Quality Estimation (QE) model for a source x and translation y be $r_{QE}(x, y)$, we build a relation $\succ_x$ as a proxy for true human preferences:

$$y_1 \succ_x y_2 \equiv r_{QE}(x, y_1) > r_{QE}(x, y_2) + \varepsilon$$

where $\varepsilon \geq 0$ is a tolerance parameter to help miti-

gate proxy model noise. We set $\varepsilon = 0.005$.

To construct preference pairs, we then select the highest scoring translation per source segment as y_w and uniformly sample a single y_l from all remaining translation candidates that satisfy $y_w \succ y_l$ under our proxy model.

This results in slightly under 8000 preference pairs (occasionally the maximum difference in COMET22 score between a segment's highest and lowest scoring sampled translations is less than ε, in which case we do not produce a preference pair), we run DPO training with a batch size of 8192 tokens (counting source, chosen and rejected tokens), a learning rate of $1e{-}6$ and $\beta = 0.5$. See Appendix 8 for a full list of hyperparameters.

At this point, we train on the preference pairs using standard DPO for 8 epochs, after which we sample a fresh set of source segments from the seed dataset, sample translations from the policy model, create a new set of preference pairs, and begin the training again. This resampling process helps ensure that the preference pairs remain relevant to the policy model during training, and leads to substantial performance improvements. In total, we perform 5 rounds of DPO training. We call this end-to-end process Direct Quality Optimization (DQO), detailed formally in Algorithm 1.

DQO can be viewed as a batched online version of DPO, as the updates are performed on batches of data sampled from the policy model.

Algorithm 1: Direct Quality Optimization

Parameters: preference relation $\succ$, number of rounds n, epochs per round m, epoch size d, learning rate α, DPO regularization β, sampled translations per source k

Input: Source language seed dataset S, reference-free QE model r_{QE}, reference model π_{ref}

$\pi_\theta \leftarrow \pi_{\text{ref}}$;
for *round* $i = 1, 2, \ldots, n$ **do**
  $X \leftarrow$ **sample** d sentences from S;
  $P \leftarrow \varnothing$;
  foreach *source* $x \in X$ **do**
    $g \leftarrow \texttt{Greedy}_{\pi_\theta}(x)$;
    $Y \leftarrow$ **sample** k translations of x from π_θ;
    $Y_+ \leftarrow Y \cup \{g\}$;
    $y_w \leftarrow \text{argmax}_{y \in Y_+} r_{QE}(x, y)$;
    $Y_l = \{y' \in Y_+ | y_w \succ_x y'\}$;
    if $Y_l \neq \varnothing$ **then**
      $y_l \leftarrow$ **sample** $y \in Y_l$;
      $P \leftarrow P \cup \{(x, y_w, y_l)\}$;
  end
  for *epoch* $j = 1, 2, \ldots, m$ **do**
    $\pi_\theta \leftarrow \texttt{DPO}(\pi_\theta, \pi_{ref}, P, \alpha, \beta)$;
  end
end

Figure 1: **Direct Quality Optimization (DQO).** $\texttt{Greedy}_\pi(x)$ is the translation of x produced with greedy search and the model π. `DPO` refers to Direct Preference Optimization – for full implementation details see Rafailov et al. (2023).

5 Experimental Results

5.1 Automatic Quality Metrics

We evaluated the model pre- and post-task alignment on the FLORES+ (Team et al., 2024) and NTREX (Federmann et al., 2022; Barrault et al., 2019) datasets, both of which cover all of the languages supported by the Megatron model.

We use corpus-level sacreBLEU[3] (Post, 2018) as well as three neural evaluation models: Reference-free CometKiwi22 (Rei et al., 2022b), reference-based COMET22 (Rei et al., 2022a), and BLEURT (Sellam et al., 2020).

Here it is important to note that the CometKiwi22 model was used as a proxy for human preferences in this experiment, and was thus directly optimized for. The scores from the other two neural evaluation models are thus more reliable measures of general model quality, and allow us to check for reward hacking, i.e. over-optimization for the CometKiwi22 model at the cost of performance.

Results are reported in Table 1 and Figure 2. We find that DPO task alignment increases all three neural quality metrics on both datasets for each of the 30 target languages. BLEU scores increased for all languages on both datasets, with the exception of Hindi, which decreased by 0.40 BLEU on NTREX and 0.4 BLEU on FLORES+ devtest, despite showing improvements on the three neural metrics, like all other languages.

Significantly, translation quality, as measured by all four translation quality metrics, improved even

[3]Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0`. For JA and ZH, we additionally use the `mecab-ja` and `mecab-zh` tokenizers.

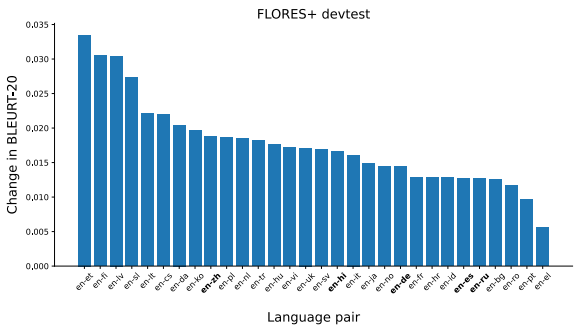


Figure 2: **Changes in BLEURT on FLORES+ devtest** with the NVIDIA Megatron EN-X model, before and after task alignment with DQO. Languages used in DQO are bolded.

for target languages unrelated to the languages used in DPO task alignment. See Appendix A.5 for the metrics for each individual language.

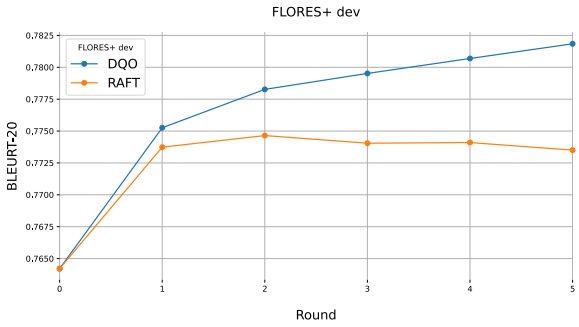


Figure 3: **Mean BLEURT-20 on FLORES+ dev at each round of DQO** with the NVIDIA Megatron EN-X model, using either Direct Quality Optimization (DQO) or Supervised Fine-Tuning (RAFT) to update the model.

To ablate the use of DPO as the update step within DQO, we perform a comparative experiment identical to DQO as described in Section 4 and Algorithm 1, but using standard supervised fine-tuning (SFT) on the preferred translation instead of DPO. Note that this is equivalent to Reward rAnked Fine-Tuning (RAFT) (Dong et al., 2023). Figure 3 shows mean performance for all language pairs through the 5 rounds of DQO. RAFT's lower performance is primarily due to catastrophic behavior for FR, JA, KO, and ZH. The poor performance of RAFT on FR, JA and KO, which were not used in RAFT training, could potentially be explained by a failure to generalize from the training data languages. Regarding ZH, which was one of the five languages used in training, we suspected unintentionally reversed labels. However, careful inspection of the training preference pairs showed no issues. See Appendix A.6, Figure 4 for charts of individual language performance and Figure 5 for mean performance after excluding the above mentioned outliers. We leave a deeper analysis to future work.

5.2 Training Data Perplexity

In order to confirm the existence of a task–data mismatch, we examine DQO's effect on model perplexity over the training data. As we do not have access to the training data used for the NVIDIA Megatron English-Many model, we repeat the above experiment with a proprietary encoder-decoder model trained on publicly available English-to-German data using the NVIDIA NeMo framework (Kuchaiev et al., 2019) (See Appendix A.4). The model architecture is similar to the Megatron model, and follows the deep encoder, shallow decoder recipe from (Kasai et al., 2021), but is larger, with a model width of 2048, a feed-forward width of 8192, 21 encoder layers, 2 decoder layers, and a 32k token vocabulary, totaling 1.3B parameters.

We apply DQO to this model as with the Megatron model, however using only English–German preference pairs. After applying DQO, we see large improvements in CometKiwi22 and COMET22 for a variety of evaluation datasets, confirming that DQO worked as expected. The arithmetic mean of perplexity over a random sample of 1 million segments from the training data increased from 7.219 (baseline model) to 9.435 (DQO), confirming that the improvements in test data preference correspond to a reduction in the model's fit to the training corpus.

5.3 Discussion

The nearly-universal improvements for both FLORES+ and NTREX in all four automatic translation quality metrics (Table 1) provide strong evidence that DQO is a suitable task-alignment algorithm for the task of producing human-preferred translations.

As shown in Section 5.2, while improving task performance, DQO increases perplexity over the training data used during supervised training. This, combined with the finding that DQO is a suitable task alignment algorithm, is evidence for the existence of the task–data mismatch.

Much of this improvement can likely be credited to general, language-agnostic changes in model behavior, even with the restriction to using only 5 of the 30 supported target languages in DQO. If task alignment of a model with a given target lan-

Source	... under the leadership of **Deng Xiaoping**.
Baseline	... tika veiktas ***Deng Xiaoping*** vadībā.
DQO	... tika veiktas **Dena Sjaopina** vadībā.
Source	... that **Carolyn Wilson** of the OHA had stolen their security deposits ...
Baseline	... ka OHA ***Carolyn Wilson*** bija nozagusi viņu drošības depozītus ...
DQO	... ka OHA darbiniece **Karolīna Vilsona** bija nozagusi viņu drošības depozītus ...
Source	... that it was **Louis Jourdain**, 16-year old son of ... **Floyd Jourdain**.
Baseline	... ka tas bija ***Louis Jourdain***, 16 gadus vecs ... **Floida *Jourdaina*** dēls.
DQO	... ka tas bija **Luiss Džordēns**, 16 gadus vecs ... **Floida Džordēna** dēls.
Source	**King Sejong** was the fourth king of the **Joseon** Dynasty ...
Baseline	***King Sejong*** bija ceturtais karalis no ***Joseon*** dinastijas ...
DQO	**Karalis Sedžons** bija ceturtais **Džosona** dinastijas karalis ...

Table 4: **Examples of translations into Latvian from the FLORES+ data set before and after DQO.** Names are bolded to highlight the DQO model's increased ability to consistently transliterate names into Latvian orthography. Names that are incorrectly transliterated are in italics. Sentences are truncated to avoid dataset leakage.

guage reduces the likelihood of untranslated source text, for instance, it would not be surprising to see similar improvements in other target languages.

Similarly, if task alignment for a given target language led to language-specific improvements (e.g., in grammar, sentence structure, punctuation, general fluency, etc.), it seems plausible that transfer learning could lead to improvements in closely related languages that have similar features.

However, manual inspection of translations before and after DQO revealed language-specific improvements in unrelated languages. In Latvian, for instance, foreign names are transliterated to match Latvian orthography and declined for grammatical case and gender, e.g. Klavinska (2021) report that *George Clooney* should be translated as *Džordžs Klūnijs*. While the baseline model applies correct transliteration occasionally and inconsistently, the DQO model almost always produces the correct transliteration. Several examples are included in Table 4.

As DQO was only performed on Chinese, German, Hindi, Russian or Spanish, none of which are closely related to Latvian, this behavior cannot have been learned from scratch during DQO. Although Chinese, Hindi, and Russian also transcribe foreign names, they use non-Latin scripts.

One possible explanation is that the baseline model learned to model both transliteration and non-transliteration, due to the range of translation quality in its supervised training data, causing inconsistent behavior at inference time. When DQO then shifts the output distribution towards certain human-preferred features, the probability of any correlated features (e.g., transliteration in Latvian), also increases.

5.4 Human Evaluation

To verify the presence of further language-specific changes for unrelated languages, we performed a human evaluation using the Multidimensional Quality Metrics framework (MQM) with professional translators (Lommel et al., 2014; Freitag et al., 2021). The translators were trained on MQM and Anthea[4], the open-source tool we used for performing MQM. We follow Freitag et al. (2021) in weighting major non-translations at 25 MQM points, other major errors at 5, and all minor errors at 1, except minor punctuation errors, which are 0.1 points.

For analysis, we selected two target languages not closely related to the languages used for task alignment: Lithuanian and Japanese.

These were selected to provide one low-to-medium resource language written in the Latin script and one in a non-Latin script, because neither is an outlier in quality metric improvement compared to the other supported language pairs, and to avoid the bias of examining Latvian, which we had already manually inspected.

For each language, we sampled complete documents (each generally two to five sentences forming a single paragraph) from FLORES+ until we had 100 source segments. The translators then annotated the baseline and task-aligned translations.

We then sorted the MQM error subcategories into two buckets, language agnostic and language specific, as seen in Table 7 in Appendix A.1.

We observe reduced error rates in both Japanese and Lithuanian in both the language-agnostic and language-specific categories (Table 5). The over-

[4] https://github.com/google-research/google-research/tree/a676d87/anthea

Language	Model	Severity				Language Specific			Weighted MQM ↓
		NT	Major	Minor	Trivial	Yes	No	N/A	
Japanese	Baseline	0	1.15	**0.61**	0.06	1.28	0.50	0.01	6.256
	DQO	0	**0.93**	0.63	**0.03**	**1.16**	**0.40**	0.01	**5.223**
Lithuanian	Baseline	0.03	0.95	0.89	0.12	1.48	0.51	0	6.402
	DQO	**0.01**	**0.80**	**0.77**	**0.10**	**1.24**	**0.44**	0	**5.030**

Table 5: **Mean number of Multidimensional Quality Metrics (MQM) errors per segment**, as annotated by professional human evaluators, with two different groupings: by severity and by whether the MQM subcategory is language specific or agnostic. NT stands for non-translation, i.e., a segment that cannot be construed as a translation of the source. Trivial refers to minor punctuation errors. This covers 100 randomly sampled English segments from the FLORES+ dataset, translated by the NVIDIA Megatron model before task alignment (baseline) and after task alignment (DQO). The weighted MQM score follows Freitag et al. (2021).

all weighted MQM score also decreased for both languages, with significant improvements in both Lithuanian ($p_u = .001$) and Japanese ($p_u = .012$), where p_u-values are conservative estimates of the true p-values computed using paired one-sided approximate randomization (Phipson and Smyth, 2010) with the Marot toolkit.[5]

5.5 DQO for Large Language Models

To compare DQO's performance against the strong baseline of other state-of-the-art DPO variants on a large language model trained specifically for translation, we apply it to the Alma-13B-LoRA model, a LLaMA-2-13B model with continued pretraining on Chinese, Czech, English, German, Icelandic, and Russian monolingual data and LoRA fine-tuning on high quality translation data (Xu et al., 2024a; Hu et al., 2022).

The highest performing human preference alignment method previously reported for this model is Contrastive Preference Optimization (CPO), a variant of DPO applied to the Alma-13B-LoRA model to create Alma-13B-R (Xu et al., 2024b). To ensure a direct comparison of optimization methods, we adopt the same data conditions and parameter masks as that prior work: restricting our seed dataset to the training data used for Alma-13B-R (the combined FLORES+ dev and devtest splits), fine-tuning only the LoRA adapters of the model, and evaluating translation out of English on the WMT'21 (for Icelandic) and WMT'22 (for the other languages) datasets.

Due to the restricted seed dataset used in this experiment, source segments are reused between rounds. As in previous experiments, we sample 8000 source segments, sample 64 translations per segment (as well as the greedy translation), and use CometKiwi22 as a proxy for human preferences. Other hyperparameters were adjusted based on a manual hyperparameter search to accommodate the differing training and sampling dynamics of LoRA training with an LLM (see Appendix A.3 for all hyperparameters).

Table 6 shows the results. The translations for ALMA-13-LoRA and ALMA-13B-R are generated with greedy inference on the publicly available model parameters[6]. This experiment indicates that DQO maintains a substantially higher BLEU score than CPO while providing similar improvements in BLEURT, COMET22, and CometKiwi22. Unlike our encoder-decoder experiments, source segments were reused between rounds to achieve a fair comparison with CPO. We would expect a higher performance with a larger pool of source data, but leave confirmation of this assumption to future work.

6 Related Work

The idea of task–data mismatch in NMT is not new. There has been extensive previous work focused on reducing this mismatch through data filtering, using surface-level heuristics (Koehn et al., 2007), statistical and neural models for alignment and quality evaluation (Sánchez-Cartagena et al., 2018; Heffernan et al., 2022; Peter et al., 2023), language identification (Lui and Baldwin, 2011; Joulin et al., 2016), or ensembles (Koehn et al., 2020).

While data filtering techniques do help reduce the task–data mismatch, they force a trade-off between increasing task alignment and retaining flawed, but potentially useful, training data. To

[5] https://github.com/google-research/google-research/tree/a676d87/marot/README.md

[6] https://huggingface.co/haoranxu/ALMA-13B-Pretrain-LoRA, https://huggingface.co/haoranxu/ALMA-13B-R

Model	English → Czech				English → German			
	BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
ALMA-13B-LoRA	79.62	88.94	83.31	**29.33**	75.06	85.14	82.19	29.65
+ DQO	80.58	89.69	**84.46**	27.72	76.03	85.95	**83.10**	**29.72**
+ CPO (ALMA-13B-R)	**80.90**	**89.73**	84.38	24.29	**76.79**	**86.24**	82.96	26.72
Model	**English → Icelandic**				**English → Russian**			
	BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
ALMA-13B-LoRA	71.64	85.32	80.84	25.06	74.25	86.90	82.55	**27.48**
+ DQO	**72.00**	85.57	**81.72**	**25.09**	75.40	87.71	**83.68**	26.68
+ CPO (ALMA-13B-R)	71.71	**86.25**	81.20	21.03	**75.74**	**88.05**	83.63	23.12
Model	**English → Chinese (simpl.)**				**Average**			
	BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
ALMA-13B-LoRA	69.79	85.54	80.56	**37.80**	74.07	86.37	81.89	**29.86**
+ DQO	**70.60**	**86.37**	**81.84**	35.58	74.92	87.06	**82.96**	28.96
+ CPO (ALMA-13B-R)	**70.60**	86.35	81.79	32.15	**75.15**	**87.32**	82.79	25.46

Table 6: Evaluation of DQO and CPO (Xu et al., 2024b) on the ALMA-13B-LoRA model. Scores are reported on the WMT'21 (Icelandic) and WMT'22 (remaining languages) test sets. The hyperparameters are specified in Appendix A.3.

counter this, curriculum learning can be used, by training first on a conservatively filtered dataset, then shifting to a cleaner subset of the data (Bogoychev et al., 2023).

However, no amount of data filtering can remove the effects of translationese, as it is present in all translations. Riley et al. (2020) and Freitag et al. (2022b) both address this by treating original and translated text as separate languages in a "multilingual" NMT model, by training either a classifier or a contrastive language model to tag each source and target segment as either original or translated. At inference time, they use their model in a zero-shot setting to translate from original source text into the distribution of original target text.

Similarly, Tomani et al. (2024) label each source sentence with a binned QE score. By adding the label of the highest quality bin to a source sentence at inference time, they successfully bias the model towards high quality translations.

Ramos et al. (2024) apply RLHF (Ziegler et al., 2020) to NMT using various QE metrics as reward, and compare it to data filtering, re-ranking using a QE model, and Minimum Bayes Risk decoding (MBR) (Kumar and Byrne, 2004; Freitag et al., 2022a), finding that a combination of data filtering, RLHF, and re-ranking performs best.

In DPO MBR fine-tuning, MBR is used to generate preference pairs for use with DPO (Yang et al., 2024). Compared to DQO, this method is computationally more expensive, and requires a reference-based QE model. In addition, DQO's online nature ensures that preference pairs remain relevant to the policy model.

Xu et al. (2024c) apply RLHF with a reward model trained to distinguish high quality references (from literary translations) and translations sampled from their model. Similar to us, they find evidence of cross-lingual transfer learning during preference learning. Specifically, when optimized only on EN–ZH, their model improved for EN to FR, ES, RU, and AR. When training only on EN–AR, however, they saw improvements in only half of the target languages.

Reward rAnked Fine-Tuning (RAFT) is the method most similar to DQO, but uses SFT to update the model towards a single preferred output rather than using DPO with a preferred/rejected output pair (Dong et al., 2023). As it was not evaluated for the translation task, used an independently trained reward model, and had slight differences in sampling parameters, we ran an ablation on whether to use DPO or SFT in DQO (see Section 4).

7 Conclusion

We demonstrate the existence of a fundamental task–data mismatch in NMT and introduce Direct Quality Optimization (DQO), a method of aligning pretrained models with human preference.

Using DQO on a multilingual NMT model, we find improvements in automatic quality metrics for all supported target languages, even those neither

used for DQO, nor related to the languages used for DQO. A human evaluation confirms that these improvements reflect increased human preference.

The improvements in translation quality for unrelated languages include language specific features that were not seen during DQO, suggesting that the baseline model had, but did not use, knowledge of those features during inference. We suggest that this is the expected behavior of a model trained with supervised learning, and present DQO as an efficient method of aligning a translation model with human preference.

In an experiment on ALMA-13B-LoRA we confirm that DQO is applicable to decoder-only LLMs.

8 Limitations

This work only tests one quality evaluation model as a proxy for human preferences, CometKiwi22, and does not examine the impact of that proxy's quality. We focused primarily on a single translation model, the NVIDIA Megatron English-Many model, using a 1.3B paramter English-German model only for the perplexity experiments (as we had access to the training data), and ALMA-13B-LoRA to verify applicability on decoder-only models. Human evaluation of translation quality was only performed on two language pairs. For all others, we relied on automatic quality evaluation metrics such as BLEURT, COMET22 and BLEU, which may not fully capture true human preference.

References

A.H. Albir. 2017. *Researching Translation Competence by PACTE Group*. Benjamins Translation Library. John Benjamins Publishing Company.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv:2204.05862.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Nikolay Bogoychev, Jelmer van der Linde, Graeme Nail, Barry Haddow, Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Lukas Weymann, Tudor Nicolae Mateiu, Jindřich Helcl, and Mikko Aulamo. 2023. OpusCleaner and OpusTrainer, open source toolkits for training machine translation and large language models. arXiv:2311.14838.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. arXiv:2304.06767.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022b. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt, Bruno Pouliquen, and Christophe Mazenc. 2016. Coppa v2.0: Corpus of parallel patent applications. building large parallel corpora with gnu make.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Antra Klavinska. 2021. Transcription of foreign personal names in the written works of learners of latvian as a foreign language. *Journal of Education Culture and Society*, 12:469–481.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In

Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, Patrice Castonguay, Mariya Popova, Jocelyn Huang, and Jonathan M. Cohen. 2019. Nemo: a toolkit for building ai applications using neural modules. arXiv:1909.09577.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Luca Sabatini, and Francesco Sassi. 2023. Translated texts under the lens: From machine translation detection to source language identification. In *Advances in Intelligent Data Analysis XXI*, pages 222–235, Cham. Springer Nature Switzerland.

Sara Laviosa. 1998. Core patterns of lexical use in a comparable corpus of english narrative prose. *Meta*, 43(4):557–570.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Yan Meng, Di Wu, and Christof Monz. 2024. How to learn in a noisy world? self-correcting the real-world data noise on machine translation. arXiv:2407.02208.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. Association for Computational Linguistics.

Belinda Phipson and Gordon K Smyth. 2010. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1).

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André F. T. Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. arXiv:2311.09132.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova,

Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez-Sánchez. 2018. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Jiajun Shen, Peng-Jen Chen, Matthew Le, Junxian He, Jiatao Gu, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2021. The source-target domain mismatch problem in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1519–1533, Online. Association for Computational Linguistics.

Jason R. Smith, Herve Saint-Amand, Magdalena Plamada, Philipp Koehn, Chris Callison-Burch, and Adam Lopez. 2013. Dirt cheap web-scale parallel text from the Common Crawl. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1374–1383, Sofia, Bulgaria. Association for Computational Linguistics.

Ilia Sominsky and Shuly Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. No Language Left Behind: Scaling neural machine translation to 200 languages. *Nature*, 630:841–846.

Brian Thompson, Mehak Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. 2024. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1763–1775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Sonja Tirkkonen-Condit. 2004. Unique items — over- or under-represented in translated language? In *Translation Universals: Do they exist?*, pages 177–184. Benjamins Translation Library.

Christian Tomani, David Vilar, Markus Freitag, Colin Cherry, Subhajit Naskar, Mara Finkelstein, Xavier Garcia, and Daniel Cremers. 2024. Quality-aware translation models: Efficient generation and quality estimation in a single model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15660–15679, Bangkok, Thailand. Association for Computational Linguistics.

Philip Williams and Barry Haddow. 2021. The elitr eca corpus. arXiv:2109.07351.

Krzysztof Wołk and Krzysztof Marasek. 2014. Building subject-aligned comparable corpora and mining it for truly parallel sentence pairs. *Procedia Technology*, 18:126–132. International workshop on Innovations in Information and Communication Science and Technology, IICST 2014, 3-5 September 2014, Warsaw, Poland.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nuo Xu, Jun Zhao, Can Zu, Sixian Li, Lu Chen, Zhihao Zhang, Rui Zheng, Shihan Dou, Wenjuan Qin, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024c. Advancing translation preference modeling with rlhf: A step towards cost-effective solution.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. arXiv:1909.08593.

A Appendix

A.1 MQM Error Subcategories by Generality

Language-agnostic	**Language-specific**	**Other**
Accuracy/Creative Reinterpretation	Fluency/Grammar	Other
Accuracy/Mistranslation	Fluency/Register	Source issue
Accuracy/Source language fragment	Fluency/Spelling	
Accuracy/Addition	Fluency/Punctuation	
Accuracy/Omission	Fluency/Character encoding	
Fluency/Inconsistency	Style/Unnatural or awkward	
Terminology/Inconsistent	Style/Bad sentence structure	
Non-translation	Terminology/Inappropriate for context	
	Locale convention/Address format	
	Locale convention/Date format	
	Locale convention/Currency format	
	Locale convention/Telephone format	
	Locale convention/Time format	
	Locale convention/Name format	

Table 7: **Multidimensional Quality Metrics error subcategories by generality**. *Language-agnostic errors* are those governed by a principle that can be generalized to all language pairs, e.g., that translations should not omit information. *Language-specific errors* are those that require additional, language-specific information to generalize from one language pair to another, e.g., correcting improper sentence structure requires knowledge of correct vs. incorrect sentence structures for a given language. *Other errors* cannot be assigned to either category.

A.2 Hyperparameters Used in Experiments on NVIDIA Megatron

Hyperparameter	Definition	Value
r_{QE}	Human preference proxy model	`CometKiwi22`
n	Number of rounds	5
m	Epochs per round	8
d	Epoch size (source sentences)	8000
α	Learning rate	1×10^{-6}
β	DPO regularization factor	0.5
k	Sampled translations per source	64
K	Top-K sampling parameter	40
P	Top-P sampling parameter	0.8
ε	Preference margin	0.005
–	Batch size	8096
–	Learning rate schedule	Linear with warmup
–	Learning rate warmup steps	150
–	Gradient clipping threshold (norm)	10

Table 8: **A list of all hyperparameters** used for Direct Quality Optimization in this paper's experiments.

A.3 Hyperparameters for experiments on ALMA-13B-LoRA

Hyperparameter	Definition	Value
r_{QE}	Human preference proxy model	`CometKiwi22`
n	Number of rounds	9
m	Epochs per round	4
d	Epoch size (source sentences)	8000
α	Learning rate	5×10^{-5}
β	DPO regularization factor	0.5
k	Sampled translations per source	64
K	Top-K sampling parameter	∞
P	Top-P sampling parameter	1.0
ε	Preference margin	0.005
–	Batch size	8096
–	Learning rate schedule	Linear with warmup
–	Learning rate warmup steps	150
–	Gradient clipping threshold (norm)	10

Table 9: **A list of all hyperparameters** used for Direct Quality Optimization in the experiments on ALMA-13B-LoRA. See https://github.com/lilt/dqo/blob/main/configs/alma-13b-lora-comparison-with-cpo-4.yaml

A.4 Composition of the DQO Seed Dataset

As described in Figure 1, Direct Quality Optimization requires a seed dataset containing input samples in the source language. This dataset does not need to include references, as the policy model π_θ is used to produce a diverse set of hypotheses, which are then scored under a QE model and transformed into preference pairs.

For our experiments, we used a general and varied seed dataset consisting of the English side of the following publicly available English–German datasets provided by the OPUS project (Tiedemann, 2012):

- bible-uedin (Christodouloupoulos and Steedman, 2015)
- CCAligned (El-Kishky et al., 2020)
- CCMatrix (Schwenk et al., 2021b; Fan et al., 2021)
- DGT v2019[7]
- EBC
- ELRA-W0143[8]
- ELRA-W0201
- ELRC-CORDIS_News[9]
- ELRC-CORDIS_Results[10]

[7] https://ec.europa.eu/jrc/en/language-technologies/dgt-translation-memory. The European Commission retains ownership of the data.

[8] https://www.elrc-share.eu

[9] https://elrc-share.eu/repository/browse/english-french-parallel-corpus-from-cordis-project-news/e4597da00ae511e9b7d400155d026706c248250ecee54d19bef388d2a42e6d93/

[10] https://elrc-share.eu/repository/browse/german-english-parallel-corpus-from-cordis-project-results-in-brief/e70e0b920ae511e9b7d400155d026706b079d7cd7f984a98ab96380f6215f358/

- ELRC-EMEA[11]
- ELRC-EU_publications[12]
- ELRC-EUR_LEX[13]
- ELRC-Information_Portal[14]
- ELRC-presscorner_covid[15]
- EMEA
- EUBookshop
- EUConst
- EuroPat[16]
- GlobalVoices
- GNOME
- JRC-Acquis v3.0 (Steinberger et al., 2006)[17]
- KDE4
- LinguaTools-WikiTitles
- MultiUN (Eisele and Chen, 2010)
- News-Commentary (Kocmi et al., 2023)
- OpenSubtitles (Lison and Tiedemann, 2016)
- ParaCrawl (Bañón et al., 2020)
- PHP
- Tatoeba
- Tilde EESC (Rozis and Skadiņš, 2017)
- TildeMODEL (Rozis and Skadiņš, 2017)
- WikiMatrix (Schwenk et al., 2021a)
- wikimedia[18]

[11] https://elrc-share.eu/repository/browse/bilingual-corpus-made-out-of-pdf-documents-from-the-european-medicines-agency-emea-httpswwwemaeuropaeu-february-2020-en-de/d6ce198a862611ea913100155d0267064011b731322946a6b897cf495fb6f023/. This dataset has been generated out of public content available through European Medicines Agency: https://www.ema.europa.eu/, in February 2020.

[12] This dataset was generated from public content available through the Publications Office of the European Union (OP Portal), https://op.europa.eu/en/home

[13] https://elrc-share.eu/repository/browse/covid-19-eur-lex-dataset-ilingual-en-mt/cf57fe82c5af11ea913100155d026706b5596d3f449a456f983bbb4e23de81a4/

[14] https://elrc-share.eu/repository/browse/information-portal-of-the-czech-president-and-czech-castle/2c11868e088b11e6b68800155d020502c402eaf049834da0bbb019049e42098c/

[15] https://elrc-share.eu/repository/browse/covid-19-eu-presscorner-v1-dataset-bilingual-en-de/67c1519c969311ea913100155d0267063c11069dcb104114901b3160c9f7618c/

[16] https://europat.net/

[17] https://joint-research-centre.ec.europa.eu/language-technology-resources/jrc-acquis_en. The European Commission retains ownership of the data.

[18] https://dumps.wikimedia.org/other/contenttranslation/

- Wikipedia (Wołk and Marasek, 2014)
- Wikititles (Kocmi et al., 2023)
- XLEnt (El-Kishky et al., 2021)

As well as the following publicly available datasets which were not obtained through OPUS:

- ELITR ECA (Williams and Haddow, 2021)
- Europarl (Koehn, 2005)
- Tilde EMA (Rozis and Skadiņš, 2017)
- Tilde RAPID 2019 (Rozis and Skadiņš, 2017)
- WIPO COPPA (Junczys-Dowmunt et al., 2016)
- WMT13 CommonCrawl (Smith et al., 2013)

These datasets were also used to train the model used in Section 5.2.

A.5 Results by Target Language

Model	Lang.	FLORES+ devtest				NTREX			
		BLEURT	COMET22	CometKiwi22	BLEU	BLEURT	COMET22	CometKiwi22	BLEU
Baseline	bg	0.8400	0.8974	0.8524	41.80	0.7713	0.8520	0.8242	32.00
DQO	bg	**0.8526**	**0.9067**	**0.8614**	**42.70**	**0.7865**	**0.8638**	**0.8341**	**32.40**
Baseline	cs	0.7758	0.8826	0.8327	32.60	0.7282	0.8509	0.8065	30.10
DQO	cs	**0.7978**	**0.9002**	**0.8504**	**34.00**	**0.7506**	**0.8696**	**0.8255**	**30.70**
Baseline	da	0.7744	0.8942	0.8396	46.40	0.7136	0.8541	0.8145	37.40
DQO	da	**0.7948**	**0.9091**	**0.8565**	**48.60**	**0.7355**	**0.8721**	**0.8341**	**39.30**
Baseline	de	0.7417	0.8535	0.8222	38.80	0.6793	0.8100	0.7950	30.80
DQO	de	**0.7561**	**0.8682**	**0.8338**	**39.30**	**0.7041**	**0.8315**	**0.8117**	**31.80**
Baseline	el	0.6738	0.8641	0.8032	25.90	0.6477	0.8494	0.7876	30.60
DQO	el	**0.6793**	**0.8699**	**0.8044**	**26.60**	**0.6567**	**0.8585**	**0.7892**	**31.60**
Baseline	es	0.7467	0.8567	0.8569	27.50	0.7304	0.8474	0.8330	40.50
DQO	es	**0.7594**	**0.8656**	**0.8662**	**28.80**	**0.7421**	**0.8547**	**0.8425**	**41.00**
Baseline	et	0.7779	0.8792	0.8421	27.10	0.7279	0.8451	0.8155	24.20
DQO	et	**0.8114**	**0.9041**	**0.8647**	**28.90**	**0.7603**	**0.8690**	**0.8399**	**25.00**
Baseline	fi	0.7959	0.8899	0.8471	24.40	0.7393	0.8550	0.8247	18.70
DQO	fi	**0.8264**	**0.9105**	**0.8640**	**26.00**	**0.7640**	**0.8736**	**0.8421**	**19.60**
Baseline	fr	0.7400	0.8638	0.8486	49.40	0.6525	0.8221	0.8289	36.10
DQO	fr	**0.7529**	**0.8713**	**0.8544**	**50.70**	**0.6632**	**0.8305**	**0.8344**	**37.00**
Baseline	hi	0.6825	0.7645	0.8040	**32.90**	0.6313	0.7227	0.7735	**25.50**
DQO	hi	**0.6991**	**0.7862**	**0.8217**	32.50	**0.6511**	**0.7459**	**0.7972**	25.10
Baseline	hr	0.8190	0.8942	0.8624	31.10	0.7707	0.8644	0.8326	31.80
DQO	hr	**0.8318**	**0.9032**	**0.8695**	**32.10**	**0.7847**	**0.8770**	**0.8445**	**32.50**
Baseline	hu	0.8378	0.8645	0.8354	26.90	0.7616	0.8141	0.8118	17.40
DQO	hu	**0.8554**	**0.8800**	**0.8488**	**27.10**	**0.7793**	**0.8294**	**0.8268**	**18.00**
Baseline	id	0.8030	0.9092	0.8414	47.50	0.7648	0.8823	0.8111	40.50
DQO	id	**0.8158**	**0.9172**	**0.8516**	**49.30**	**0.7784**	**0.8917**	**0.8251**	**41.10**
Baseline	it	0.7699	0.8725	0.8590	30.60	0.7280	0.8455	0.8279	36.70
DQO	it	**0.7860**	**0.8821**	**0.8676**	**31.40**	**0.7467**	**0.8613**	**0.8434**	**37.50**
Baseline	ja	0.6832	0.8918	0.8545	32.60	0.6042	0.8584	0.8251	26.40
DQO	ja	**0.6981**	**0.9019**	**0.8629**	**34.10**	**0.6208**	**0.8713**	**0.8395**	**27.10**
Baseline	ko	0.6538	0.8689	0.8433	29.40	0.5788	0.8317	0.8085	25.50
DQO	ko	**0.6734**	**0.8820**	**0.8550**	**30.30**	**0.5980**	**0.8481**	**0.8250**	**26.50**
Baseline	lt	0.8043	0.8742	0.8344	27.30	0.7485	0.8404	0.8057	21.60
DQO	lt	**0.8264**	**0.8910**	**0.8490**	**28.80**	**0.7699**	**0.8564**	**0.8181**	**22.30**
Baseline	lv	0.7896	0.8677	0.8253	30.50	0.6997	0.8097	0.7816	20.40
DQO	lv	**0.8201**	**0.8902**	**0.8431**	**32.10**	**0.7418**	**0.8424**	**0.8088**	**21.70**
Baseline	nl	0.7425	0.8617	0.8483	27.00	0.7080	0.8384	0.8205	34.20
DQO	nl	**0.7611**	**0.8756**	**0.8601**	**28.10**	**0.7262**	**0.8556**	**0.8356**	**35.40**
Baseline	no	0.7771	0.8899	0.8526	33.80	0.7447	0.8622	0.8267	36.90
DQO	no	**0.7915**	**0.8991**	**0.8646**	**34.00**	**0.7644**	**0.8779**	**0.8445**	**38.70**
Baseline	pl	0.7600	0.8678	0.8206	21.40	0.6992	0.8312	0.7939	25.70
DQO	pl	**0.7787**	**0.8818**	**0.8312**	**22.80**	**0.7153**	**0.8463**	**0.8058**	**26.80**
Baseline	pt	0.7856	0.8941	0.8453	50.80	0.7069	0.8477	0.8236	33.90
DQO	pt	**0.7952**	**0.9000**	**0.8531**	**51.20**	**0.7197**	**0.8574**	**0.8341**	**35.00**
Baseline	ro	0.8026	0.8927	0.8594	40.30	0.7338	0.8441	0.8255	33.30
DQO	ro	**0.8144**	**0.9015**	**0.8645**	**41.40**	**0.7474**	**0.8571**	**0.8386**	**34.70**
Baseline	ru	0.7430	0.8755	0.8329	31.30	0.6706	0.8299	0.8002	31.80
DQO	ru	**0.7556**	**0.8842**	**0.8419**	**32.00**	**0.6831**	**0.8433**	**0.8104**	**31.90**
Baseline	sl	0.7978	0.8679	0.8359	30.00	0.7174	0.8106	0.7877	28.30
DQO	sl	**0.8252**	**0.8860**	**0.8517**	**31.80**	**0.7576**	**0.8410**	**0.8163**	**29.60**
Baseline	sv	0.7945	0.8957	0.8515	45.40	0.7401	0.8581	0.8192	40.90
DQO	sv	**0.8113**	**0.9064**	**0.8650**	**46.20**	**0.7632**	**0.8781**	**0.8400**	**42.40**
Baseline	tr	0.7693	0.8827	0.8441	29.10	0.6802	0.8235	0.8129	17.60
DQO	tr	**0.7875**	**0.8953**	**0.8559**	**30.10**	**0.7011**	**0.8402**	**0.8287**	**17.70**
Baseline	uk	0.7432	0.8728	0.8172	29.80	0.6678	0.8230	0.7838	24.80
DQO	uk	**0.7603**	**0.8878**	**0.8300**	**30.50**	**0.6868**	**0.8423**	**0.7983**	**25.80**
Baseline	vi	0.7157	0.8736	0.8299	42.20	0.6753	0.8442	0.8081	41.30
DQO	vi	**0.7329**	**0.8857**	**0.8429**	**43.80**	**0.6917**	**0.8589**	**0.8234**	**42.10**
Baseline	zh	0.7015	0.8582	0.8199	42.00	0.6267	0.8099	0.7879	34.50
DQO	zh	**0.7202**	**0.8752**	**0.8367**	**44.10**	**0.6468**	**0.8292**	**0.8067**	**36.00**

Table 10: Automatic quality evaluation metrics for all target languages supported by the NVIDIA Megatron model, before and after Direct Quality Optimization (DQO), computed on both the FLORES+ devtest and NTREX datasets.

A.6 Ablation of Update Step: DPO vs. SFT

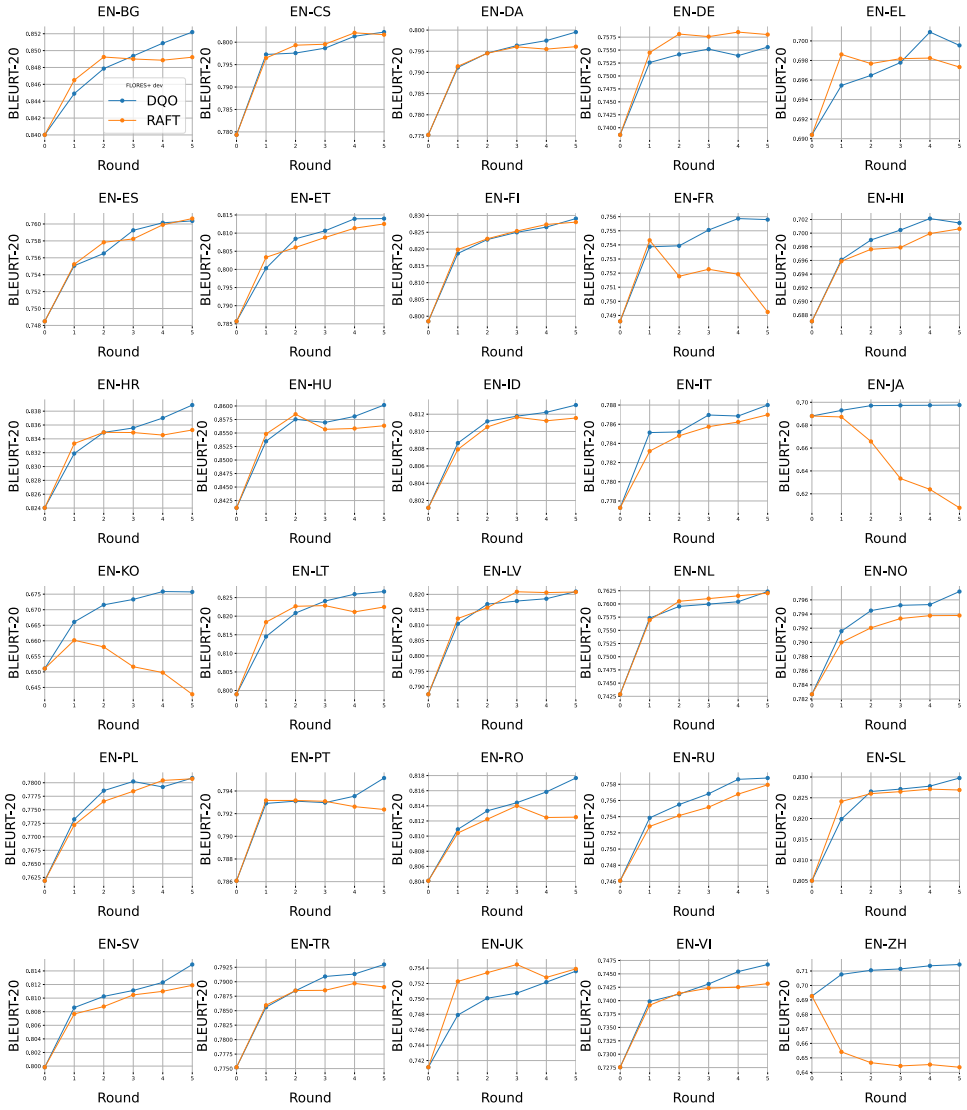


Figure 4: **Mean BLEURT-20 per language pair on FLORES+ dev after each round of DQO** with the NVIDIA Megatron EN-X model, using either Direct Preference Optimization (DPO) or Supervised Fine-Tuning (SFT) to update the model. DQO with SFT is equivalent to Reward rAnked Fine-Tuning (RAFT).

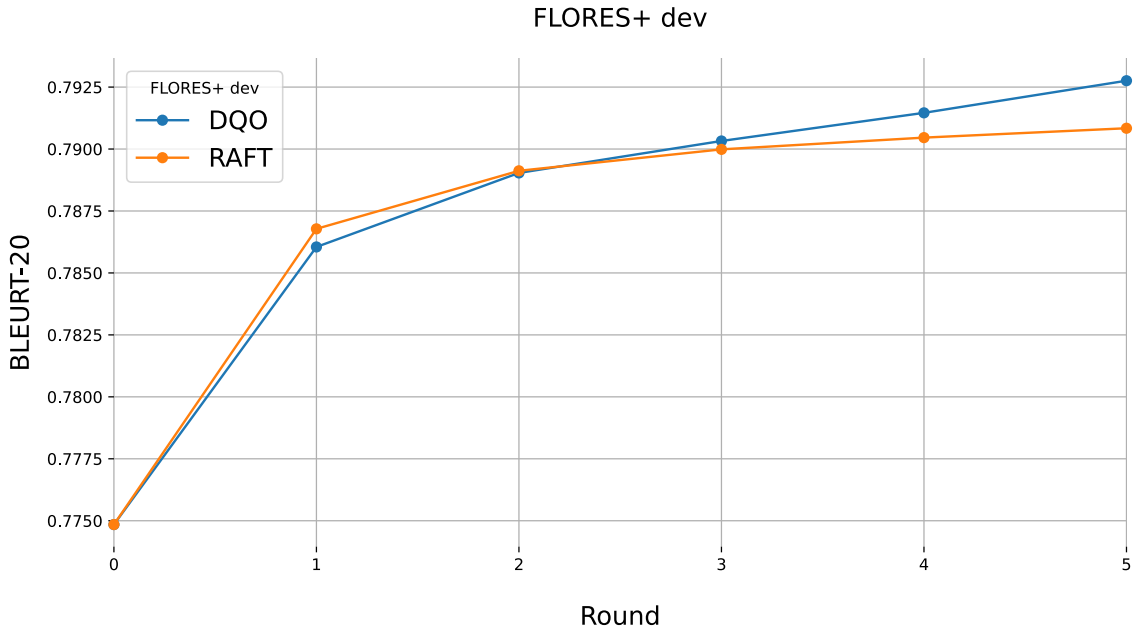


Figure 5: **Mean BLEURT-20 on FLORES+ dev, excluding outliers after each round of DQO** with the NVIDIA Megatron EN-X model, using either Direct Preference Optimization (DPO) or Supervised Fine-Tuning (SFT) to update the model. DQO with SFT is equivalent to Reward rAnked Fine-Tuning (RAFT). English to French, Chinese, Japanese, and Korean were excluded from this chart as outliers. See Figure 3 for the chart including outliers.

Audio-based Crowd-sourced Evaluation of Machine Translation Quality

Sami Ul Haq[1,2], **Sheila Castilho**[1,2], **Yvette Graham**[2,3]
[1]ADAPT Centre
[2]Dublin City University (DCU), Ireland
[3]Trinity College Dublin (TCD), Ireland
sami.haq2@mail.dcu.ie sheila.castilho@dcu.ie ygraham@tcd.ie

Abstract

Machine Translation (MT) has achieved remarkable performance in recent times, with growing interest in speech translation and multimodal approaches. However, despite these advancements, MT quality assessment remains largely text-centric, typically relying on human experts who read and compare texts. Since many real-world MT applications (e.g., Google Translate Voice Mode, iFLYTEK Translator) involve translation being spoken rather than printed or read, a more natural way to assess translation quality would be through speech as opposed text-only evaluations. This study compares text-only and audio-based evaluations of 10 MT systems from the WMT General MT Shared Task, using crowd-sourced judgments collected via Amazon Mechanical Turk. We additionally, performed statistical significance testing and self-replication experiments to test reliability and consistency of the proposed audio-based approach. Crowd-sourced assessments based on audio yield rankings largely consistent with text-only evaluations but, in some cases, identify significant differences between translation systems. We attribute this to the richer, more natural modality of speech and propose incorporating speech-based assessments into future MT evaluation frameworks.

1 Introduction

Reliable evaluation process is critical in the development and refinement of MT systems. MT evaluation (MTE) often relies on both automated and manual measurement techniques. Manual evaluation is always a preferred choice and provides a deeper understanding of system quality, while automatic evaluation metrics (AEMs) often serve as a proxy for human judgment (Castilho et al., 2018). AEMs support reusable assessments, system comparison and rapid MT deployment. However, AEMs face several issues including their inability to handle contextual and cultural nuance, the dependency on reference translation, and domain-specific challenges. Therefore, despite being time-consuming and expensive, human assessment is still a fundamental requirement for reliable evaluation.

The annual Conference on Machine Translation (WMT) is the primary forum for collecting human judgments to evaluate metrics and participating systems in its shared translation task each year. In early evaluation campaigns, 5-point adequacy and fluency ratings were gathered from participants as the primary evaluation metric (Koehn and Monz, 2006). Subsequent WMT campaigns adopted a ranking-based evaluation approach as the official metric (Vilar et al., 2007), with rankings still collected from participants of the evaluation campaign. Regarding fluency as a measure of MT output quality, Graham et al. (2013a) argued that using a 1-100 continuous scale yields better inter-annotator consistency compared to a five-point interval scale. Supporting this, Bojar et al. (2016) found strong correlations between adequacy and fluency-based evaluations. These findings led WMT to replace relative ranking with adequacy-based Direct Assessment (DA) on a continuous scale as the official metric (Bojar et al., 2017). For into-English translation tasks, WMT frequently relied on crowd-workers for its human evaluation campaigns. Crowd-based evaluations allow for a fast and cheap MT quality evaluations (Callison-Burch, 2009). When coupled with quality-controlled annotations, non-expert crowd assessments show better inter-annotator consistency (Graham et al., 2013a, 2017). However, Castilho et al. (2017b) found that crowd-workers, compared to professional translators, were less capable of detecting subtle MT errors. Studies by Läubli et al. (2018) and Toral et al. (2018) also favored the use of professional translators over researchers or crowd-workers due to their ability to differentiate between human and machine translations. Consequently, WMT revised its evaluation procedures to prioritize professional transla-

Proceedings of the Tenth Conference on Machine Translation, pages 52–63

tors over crowd-workers (Kocmi et al., 2022, 2023). Despite its limitations, crowd-based assessment remains the most convenient choice for certain tasks, particularly monolingual DA, which does not require human raters to have bilingual knowledge (Graham et al., 2017), making it easier to conduct. More recently, WMT performed evaluations using Error Span Annotation (ESA) protocol (Kocmi et al., 2024), which requires annotators to assign an overall score to each segment, similar to DA and classify errors based on severity (e.g. major or minor).

The human evaluation process has evolved over time; however, there is still no consensus on the best approach to evaluating translation quality (Castilho et al., 2018). Current MT evaluation metrics primarily considers text, despite the fact that many real-world MT applications involve spoken rather than written translation. Most importantly, the recent emergence of pre-trained multimodal models (Barrault et al., 2023) has enabled support for direct speech-to-speech, text-to-speech and speech-to-text translation, however appropriate methods for evaluation for these systems are yet limited or borrowed from text-domain (Salesky et al., 2021; Sperber et al., 2024).

We argue that speech, as a natural and expressive modality, can provide more reliable measures of MT quality. To support this claim, we propose incorporating text-to-speech (TTS) technology into direct MT assessment, allowing for a direct comparison between text-only and speech-enabled evaluation approaches. Our study collects human judgments for German-English translations from WMT shared task using crowd-workers hired via Amazon Mechanical Turk. The evaluation consists of two conditions: (i) a text-only setup, replicating the conventional method where evaluators compare written MT output with a reference translation, and (ii) a text-audio setup, where evaluators listen to the MT output while reading the reference translation. We perform self-replication experiments and statistical significance tests to assess the consistency and reliability of the proposed method.

A comparative analysis of these evaluation conditions yields two key findings. First, rankings derived from text-audio evaluations are broadly similar to the original evaluations but also show notable differences compared to conventional setups, with the audio-based method demonstrating a substantially greater ability to detect significant differences between translation systems. We hypothesize that this difference arises because speech is a natural and rich modality, capable of conveying prosodic and expressive features that text alone cannot capture. Second, consistent with prior research, our results confirm that crowd-workers tend to assign lower rankings to human translations that diverge from the reference, while favoring literal machine translations (Castilho et al., 2017a; Fomicheva, 2017). Furthermore, self-replication experiments reveal a higher positive correlation between repeated runs of audio-based evaluations, indicating improved reliability and consistency of this new approach.

Domain	#segments	Avg. doc length
conversation	462	6.8
ecommerce	501	18.5
news	506	14.5
social	515	15.6

Table 1: Number of segments and average document length (#segments per document) of German-English data used in the general translation test sets.

2 Methodology

2.1 Data set

We used MT outputs from WMT 2022 German-English translation task, comprising around 20,000 translations submitted by 10 participating systems, with each system contributing approximately 2,000 translations. This original evaluation set is a bilingual corpus drawn from different domains, as shown in Table 1, with document lengths varying considerably by domain. To ensure balanced domain representation while preserving document order, a subset of documents was randomly sampled from each domain. We use on average 450 segments per system for multimodal[1] and text-only experiments.

The WMT evaluation campaign has already published results from crowd-based human evaluations of the submitted systems. As WMT now conducts bilingual ('source-based') evaluations using professional translators, we focus on WMT 2022—the most recent workshop to perform monolingual DAs.

[1]In this study, multimodal is used to refer to text-audio based setup

2.2 Assessment Design

AMT crowd-sourcing service was used to design and collect human judgments, with each task consisting of 100 segments. A single segment along with a reference translation is presented at one time. Where possible, segments are collected and shown in document context. In adequacy based assessments, crowd-workers are asked to rate how adequately an MT output expresses the meaning of the reference translation. The scores are collected on 0–100 visual analog scale (VAS) for each segment. Additionally, rater quality control mechanism is implemented to filter out ratings from non-reliable raters, as outlined by Graham et al. (2017). At the end of the task, evaluators have the option to provide feedback on their experience.

The segment-level ratings were used to calculate system-level rankings. At the end of the evaluation, we provide two types of segment-level scores, averaged across one or more raters: raw scores and z-scores, with the latter standardized for each annotator. The final score of an MT system is the mean standardized score of its ratings after filtration. Multiple judgments are collected per segment, increasing the number of annotators per translation enhances the consistency and reliability of the mean score. Since reference-based assessment required only knowledge of the English language; the selection criteria required participants to be native English speakers.

We compare judgments collected using following two different setups:

- *Text-only*: MT output and reference translations, both are presented as text (Figure 1).
- *Multimodal*: MT output is presented in audio (TTS) and reference translation as text (Figure 2).

Overall, we gathered approximately 12,000 crowd-sourced judgments for German-English language pair using DA. Compared to ordinal ranking or relative preference judgments (Callison-Burch, 2009), direct estimation facilitates more robust statistical analysis, thus making it suitable for crowd-sourced annotations (Graham et al., 2013a). When combined with quality control mechanisms, direct assessments have shown effective and relatively consistent human judgments of MT quality in WMT evaluation campaigns (Specia et al., 2020; Akhbardeh et al., 2021; Kocmi et al., 2022).

2.2.1 Text-only setup

We randomly sampled 500 segments per system (with the addition of quality control segments, the total could be increased). The selected translations are then converted into bit-mapped images, in order to deter workers from using speech feature of Web Browsers to read-aloud the translations.

In this scenario, the workers are shown the reference and the MT output as text and asked to rate MT output by moving the slider (as shown in Figure 1). For task simplicity, we kept the structure of assessment similar to existing evaluation setups (Graham et al., 2017; Kocmi et al., 2022). The ratings are collected per segment in a sequential manner, adhering to the document order where feasible. However, longer documents may need to be divided into smaller units to comply with the limit of 100 segments per task. The setup restricts assessors from revisiting and modifying ratings of previous segments to ensure integrity of quality control measures.

2.2.2 Multimodal setup

For comparison, the same segments sampled for the text-only scenario were considered in this experiment. However, this setup utilises TTS technology to present the MT output in an audio-equivalent form. To make the task less cognitively taxing, we present only the MT system's output in audio form. For this, we used the Google Cloud Text-to-Speech (TTS) Service (GCS)[2] to generate audio representations of MT outputs. The service was employed with its default human-like voice settings, which are noted for their high quality and clarity. GCS is well-suited for long-form content[3] due to its close approximation of human speech and its ability to provide an enhanced listening experience (Cambre et al., 2020).

2.2.3 HITs

Both multimodal and text only assessments are carried out separately. Each task, referred to as "HITs" (Human Intelligence Task) contains 100 translations in total for each setup. In addition to system output, a set of quality control segments was added, keeping the total size of HIT to 100. The quality control segments consists of exact repeats (*ask_again*) and degraded translations

[2] https://cloud.google.com/text-to-speech

[3] For multimodal experiments, in total a human assessor may have to listen up-to 20 minutes of machine translation outputs, therefore along with accuracy of TTS, a pleasant listening experience is important.

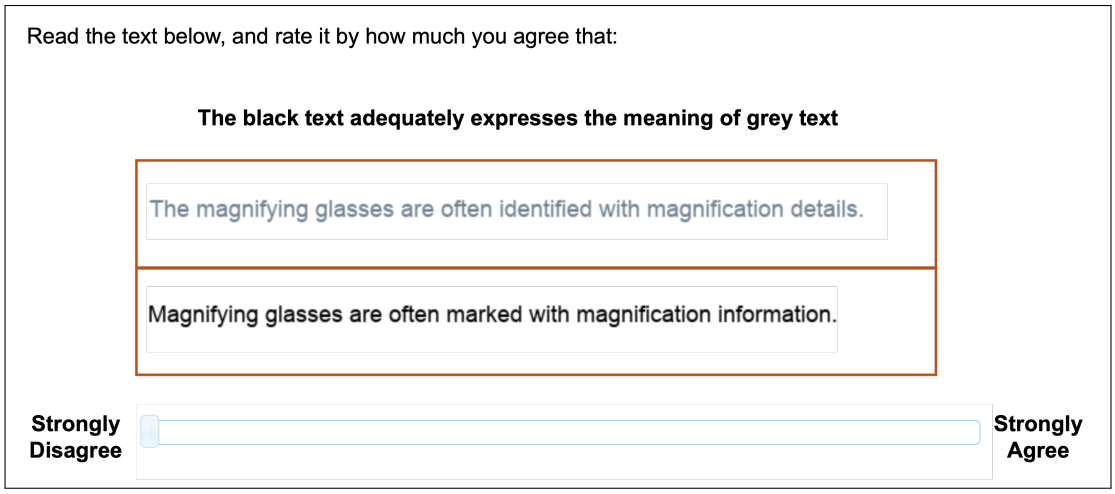


Figure 1: Screenshot of the text-only assessment interface, as presented to an AMT worker. Reference text is presented in grey while MT output is shown in black text. The slider is initially placed at left most corner; workers move it to the right in reaction to the question.

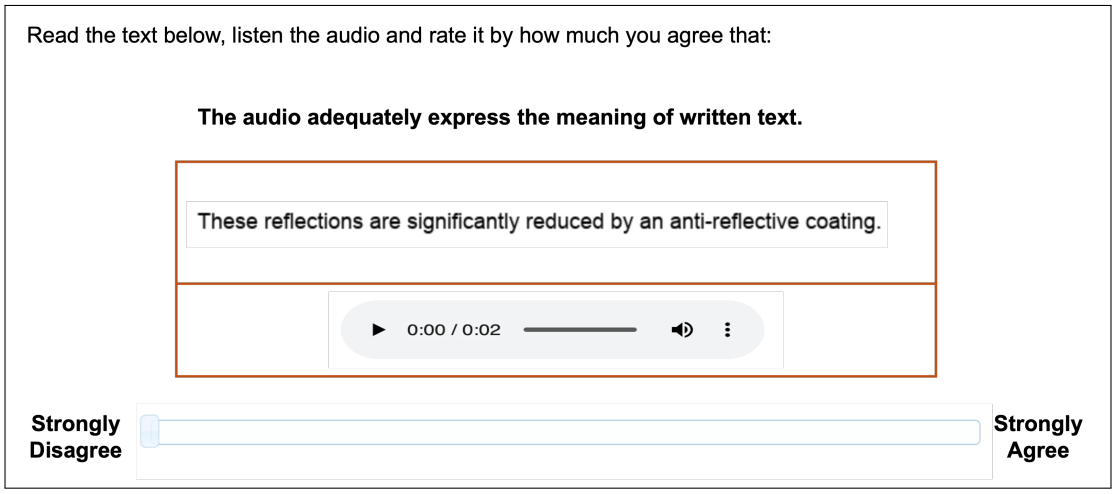


Figure 2: Screenshot of the multimodal assessment interface, as presented to an AMT worker. Worker can use audio control to listen translations, the text in presented in the image form. The slider is initially placed at left most corner; workers move it to the right in reaction to the question.

(*bad_reference*), duplicated from system outputs. Thus, each HIT consists of approximately 20% quality control segments (used to estimate workers' reliability) and 80% genuine system outputs. To create *bad_reference* pairs, we followed the strategy of randomly substituting words in a sentence, as outlined in Graham et al. (2013b). For the multimodal setup, the quality control segments were first prepared using the same strategy in text form and then converted into audio using the TTS API.

Judgments from crowd workers with limited or no knowledge of the assigned task pose a significant risk of inconsistency and discrepancies in the results. Expert-based MT quality assessment is the preferred approach; however, it incurs high economic and time costs, making crowdsourcing a viable alternative. Consequently, assessing worker reliability becomes critically important in crowd-sourced evaluations. Quality control segments within HIT allow for reliability estimates based on workers distribution of scores assigned to *bad_reference* and *ask_again* items. These estimates are based on following two assumptions:

1. The consistent assessor will assign significantly higher score to the system producing high quality translations compared to a system producing inferior outputs.

2. The consistent assessor will assign highly similar scores in repeated evaluations of the same translations.

Analysis of assumptions 1 and 2 can provide a measure of workers' ability to differentiate between a good and inferior translation. Assumptions 1 and 2, based on the sets of *bad_reference* and *ask_again* translations, posit that a consistent worker would assign significantly lower scores to degraded (*bad_reference*) translations and similar scores to repeated (*ask_again*) translations. For this, we apply the Wilcoxon rank-sum test to compare the score differences between *ask_again* and *bad_reference* translation pairs, with a resulting p value as an estimate of reliability. The expectation is that the difference in scores for degraded translation pairs will be smaller than for repeated judgments. A lower p-value ($p < 0.05$) indicates higher reliability, demonstrating that the worker can effectively distinguish between high-quality and degraded translations. As shown in Figure 3, conscientious workers assigned lower scores to degraded translations compared to the original references. Furthermore, for repeated segments, they exhibited a consistent scoring pattern by assigning similar scores to identical pairs.

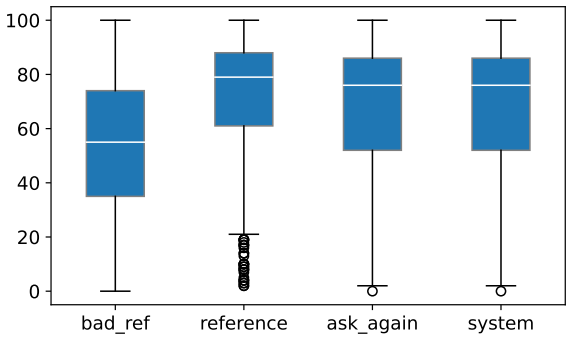


Figure 3: Score distributions of reliable workers across different quality control segments: $bad\_reference$, ask_again, and original system outputs.

Table 2 provides statistics on the number of workers involved in each assessment type and the percentage of workers who passed the quality control threshold. A similar trend was observed across both assessment types, with nearly 20% of workers meeting the reliability criteria. To determine whether to accept or reject HITs, the mean score differences for *bad_reference* and *ask_again* pairs were carefully analyzed, rather than relying solely on automatic quality control checks.[4] This is further reflected in the difference between the number of approved workers and those who met the quality control criteria. Rejected HITs were rescheduled to obtain fresh judgments.

[4]In addition to statistical tests, other measures were in place to detect robotic or low-quality submissions, such as extremely short completion times, lack of slider movement, and assigning the same rating to every judgment.

3 Results

System rankings are calculated for each setup using filtered judgments—only those that passed the quality control criteria. Quality control segments (*bad_reference* and *ask_again*) are excluded from the final system rankings. System rankings are based on the mean raw and standardized (z) scores. To compute the standardized score for each system, individual scores are first normalized using each worker's mean and standard deviation (as per equation 1). The standardized scores for all segments corresponding to a system are then averaged to obtain the system-level score (Graham et al., 2014). Since HITs are structured so that a single worker may assess multiple systems, standardising the scores helps mitigate individual biases and harmonise outputs across workers.

Table 3 presents the raw and standardized scores of the participating systems across different experiments, with the last three columns showing the official results from WMT. Systems are ordered from best to worst based on their average standardized scores, with the raw score used as a secondary criterion when standardized scores are identical to two decimal places.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

The text-only results show increased correlation between the raw and standardised score, with few exceptions such as system `LT22` and `JDExploreAcademy`, which would have ranked better according to raw score. This close correlation suggests an even distribution of segments from different systems across workers and may also be attributed to the homogeneous nature of the task (text-only data). It is important to note that these rankings may not fully reflect actual system performance or align with the official WMT rankings, as we used a smaller set of judgments per system compared to WMT22, with the primary objective of investigating and comparing audio-based and text-based evaluations.

In the multimodal scenario, the differences in system rankings between z scores and raw scores are more pronounced. Based on the raw scores, a

	Workers			Translations		
modality	Total	Approved	Pass QC	Total	Approved	Pass QC
text only	225	47	42 (18.50%)	23.3k	5.1k	4.6k (19.7%)
multimodal	242	52	48 (19.83%)	26.1k	6.0k	5.3k (20.3%)

Table 2: Numbers of workers and translations, before and after quality control for multimodal and text only experiments.

different ranking emerges, with `Lan-Bridge` performing best. This divergence may be caused by the differing nature of the evaluation setup, particularly the use of both audio and text for evaluation. The out-of-sequence numbers in order column (Table 3) highlight differences in system rankings across different experiments.

A direct comparison of the mean standardized scores across both tables reveals substantial differences in system rankings. For example, in the text-only evaluation, `Online-W` outperforms `PROMT` based on standardized scores, whereas the multimodal evaluation ranks `PROMT` as the top-performing system. Similarly, `Online-G` is ranked sixth, below `Online-A`, `Online-W`, and `Online-Y` in the text-only setup, but is rated higher than these systems in the multimodal evaluation. For most other systems, rankings diverge by one or two places between setups, with the exception of `Human-B`, which consistently ranks as the lowest-performing system in both evaluations. Ideally, `Human-B` (the human reference translation) should be the top-performing system. However, the results suggest that crowd-workers struggled to distinguish between human translations and MT outputs. This aligns with prior research suggesting that crowd-workers tend to favor literal, straightforward translations, resulting in lower rankings for human translations that deviate from the reference (Fomicheva, 2017; Freitag et al., 2021).

3.1 Significance Test Results

Since both approaches yield different system rankings without a clear indication of which better reflects actual performance, more robust testing is required to determine whether the observed ranking differences are genuine. To address this, we employ two techniques: statistical significance testing and self-replication. Significance testing estimates the likelihood that ranking differences between system pairs occurred by chance, while self-replication examines the reproducibility of results to verify their reliability and consistency.

The results of significance tests are visualised as heat maps in Figure 4 for the multimodal and text-only setups. Specifically, we apply one-sided Wilcoxon rank-sum test to compare the standardized human assessment score distributions for each pair of systems.

Tables with head-to-head comparisons between all systems are included in Appendix A.

The significance matrices are constructed under the hypothesis that the scores of system X are significantly better than those of system Y at a given confidence level, p. A comparison of the text-only and multimodal heat maps reveals that the multimodal approach results in a slightly higher proportion of significant differences between systems with fewer uncertainties. For example, at a confidence level of $p < 0.05$, the text-only method identifies relatively few significant differences, whereas the multimodal method demonstrates more distinct separations among systems. For example, the multimodal heat map shows that `Online-G` performs significantly better than `JDExploreAcademy`, `LT22`, `Lan-Bridge`, and `Online-B`, as confirmed by its higher multimodal average z-score. Similarly, for `Online-A`, both the text-only and multimodal evaluations lead to similar conclusions.

3.2 Self-replication Results

Figure 5 presents scatter plots comparing initial and self-replicated judgments from multimodal and text-only experiments. To assess the consistency of judgments collected using the multimodal (text and audio) approach, we conduct two independent runs and compute the Pearson correlation (r) between the initial and self-replicated results. In Figure 5 (a), self-replicated and original multimodal assessments are plotted on the x-axis and y-axis, respectively. Figure 5 (b) illustrates the correlation for text-only and multimodal scores, with the former on the x-axis and the latter on the y-axis. A high correlation would be indicated by points

System	text			official-text			multimodal		
	raw ave.	ave. z	order	raw ave.	ave. z	order	raw ave.	ave. z	order
PROMT	73.05	0.14	3	66.02	-0.127	10	69.63	0.19	1
Online-G	68.81	0.06	6	64.1	-0.057	4	68.76	0.19	2
Online-A	74.06	0.19	2	67.3	-0.070	5	74.61	0.18	3
Online-W	74.16	0.22	1	70.8	-0.023	2	70.74	0.14	4
Online-Y	72.16	0.13	5	66.5	-0.089	7	69.89	0.14	5
Online-B	71.49	0.14	4	66.3	-0.092	8	67.10	0.08	6
JDExploreAcademy	72.89	0.05	8	68.1	-0.038	3	67.85	0.07	7
LT22	74.36	0.05	7	64.8	-0.126	9	64.52	0.07	8
Lan-Bridge	68.29	0.05	9	68.8	0.004	1	71.24	0.04	9
Human-B	66.94	-0.12	10	68.3	-0.086	6	63.45	-0.16	10

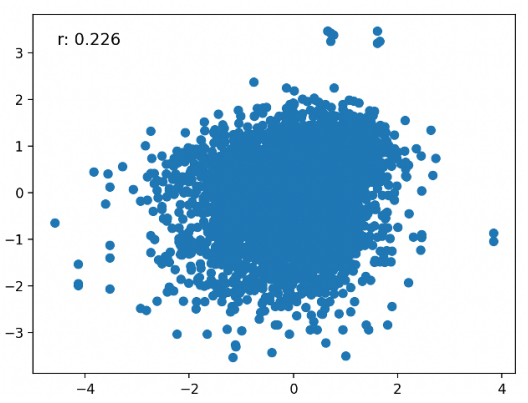
Table 3: Comparison and system rankings based on scores from the text-only and multimodal (text + audio) setup for the German–English translation direction. Systems are ordered by their average standardized (z) scores. In cases of a tie in z scores, the average raw (raw ave.) score is used as a secondary ranking criterion.

Figure 4: Significance test outcomes for text-only and multimodal method of human evaluation. Colored cells indicate that the scores of the row i system are significantly greater than those of the column j system.

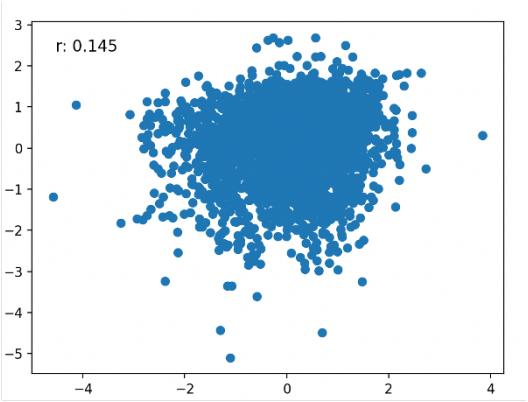
closely aligning with a straight line. While both approaches show a weak positive correlation, the multimodal setup exhibits a slightly higher correlation than the text-only setup, suggesting the potential of audio-based evaluation for providing reliable MT quality estimates.

3.3 Discussion

The results of significance and correlation tests suggest that speech can offer consistent and valuable insights into MT quality. We hypothesize that these differences arise because speech is a richer modality, capable of conveying prosodic and expressive features (Kraut et al., 1992). As a result, evaluators listening to translations were better able to detect major variations and unnatural-sounding MT outputs. Furthermore, feedback from evaluators at the end of the assessment indicated no challenges with audio-based evaluation. For instance, one worker stated that "all the audio samples were good," while another noted that "the audio is very clear". A general comment read, "the HIT is very unique, and there were no issues during the experiment". These preliminary results, obtained using non-expert crowd workers, suggest the effectiveness of speech in MT evaluation. However, further investigation may be required, and a more fine-grained approach—such as error annotation—could help better quantify the impact of audio in MT assessment.



(a) Self-replication (multimodal vs multimodal)



(b) text-only vs multimodal

Figure 5: Scatter plots illustrating the correlation (r) between different evaluation approaches. (a) shows the correlation between two runs of the multimodal approach, while (b) compares the results of the text-only and multimodal approaches.

4 Conclusion

We have presented our findings on integrating speech into human evaluation of MT quality. Our experiments with crowd workers compared MT system rankings from text-only and speech-enabled evaluation setups.

Despite using basic TTS tools and crowd workers, our study extends MT evaluation beyond traditional text-based assessments, highlighting the potential of audio-based evaluation to provide distinct insights into MT evaluation. As MT research increasingly embraces multimodal translation, our findings provide empirical evidence that text-only evaluation may be insufficient. Beyond MT, this approach could benefit fields such as automatic dubbing, AI-assisted interpreting, and multilingual speech interfaces. Overall, our study emphasizes the need for more holistic evaluation benchmarks that better reflect the complexity of real-world language use.

Our code[5] and collected human annotation data are freely available.

5 Limitations

We performed a general adequacy-based MT evaluation using crowd-workers on a limited dataset. Since the primary goal was to test whether audio-based judgments make a difference, we employed a simplified assessment approach and focused only on the German-English language pair. We acknowledge that even expert-based human judgments can be noisy, potentially leading to low inter-annotator agreement (IAA) if not carefully conducted. Nevertheless, we collected a large sample of annotations from crowd-workers to compare the two approaches. With intrinsic quality control measures, crowd-sourced annotations have been shown to achieve higher IAA (Graham et al., 2017). However, due to limited time and platform constraints, manual filtering of noisy annotations was not feasible, making it difficult to eliminate all low-quality responses. Furthermore, we did not calculate inter-annotator or intra-annotator agreement, as these aspects have already been extensively studied in the context of crowd-sourced direct assessment (Graham et al., 2013a, 2017).

Regarding the TTS model, we relied on a single vendor and did not conduct comparisons across different voices, speech rates, or providers. This restricts the generalizability of our findings, as results may vary with alternative TTS configurations. Nonetheless, we selected the model judged to have the most human-like voice based on a review of the vendor's technical documentation.

As MT quality evaluation has increasingly moved toward ESA-style (Kocmi et al., 2024) annotations (at least in WMT), audio-based evaluation could be integrated into such platforms to identify error spans by listening to translations and assigning final scores. However, accurately segmenting the audio for this purpose would pose a significant challenge.

6 Ethical Considerations

The human annotations collected via Amazon Mechanical Turk were fully anonymous. Anonymous users with MTurk accounts (meeting the defined criteria) submitted the tasks using numeric worker IDs. Although no personal identity information

[5] https://github.com/sami-haq99/Multimodal_Direct_Assessment

was revealed, we removed the worker IDs after payments were processed. Since the crowd-workers only needed to be native speakers and were not required to be expert translators, they were compensated according to the platform's minimum task rate.

In cases where crowd-workers did not meet the quality control criteria—such as submitting robotic responses or completing tasks in an unrealistically short time—we rejected their submissions and did not provide payment.

Acknowledgements

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

The Authors also benefit from being members of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017. Findings of the 2017 conference on machine translation (wmt17). Association for Computational Linguistics.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, et al. 2016. Findings of the 2016 conference on machine translation (wmt16). In *First conference on machine translation*, pages 131–198. Association for Computational Linguistics.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 286–295.

Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of Voices: A Large-Scale Evaluation of Text-to-Speech Voice Quality for Long-Form Content. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Honolulu HI USA. ACM.

Sheila Castilho, Stephen Doherty, Federico Gaspari, and Joss Moorkens. 2018. Approaches to human and machine Translation Quality Assessment. In *Translation Quality Assessment: From Principles to Practice*, volume 1 of *Machine Translation: Technologies and Applications*, pages 9–38. Springer International Publishing.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Panayota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio Miceli-Barone, and Maria Gialama. 2017a. A comparative quality evaluation of PBSMT and NMT using professional translators. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 116–131, Nagoya Japan.

Sheila Castilho, Joss Moorkens, Federico Gaspari, Andy Way, Panayota Georgakopoulou, Maria Gialama, Vilelmini Sosoni, and Rico Sennrich. 2017b. Crowdsourcing for nmt evaluation: Professional translators versus the crowd. *Translating and the Computer*, 39.

Marina Fomicheva. 2017. *The Role of human reference translation in machine translation evaluation*. Ph.D. Thesis, Universitat Pompeu Fabra. Accepted: 2017-08-01T10:07:21Z Publication Title: TDX (Tesis Doctorals en Xarxa).

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. Place: Cambridge, MA Publisher: MIT Press.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013a. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013b. Crowd-Sourcing of Human Judgments of Machine Translation Fluency. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 16–24, Brisbane, Australia.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the workshop on statistical machine translation*, pages 102–121. Association for Computational Linguistics.

Robert Kraut, Jolene Galegher, Robert Fish, and Barbara Chalfonte. 1992. Task requirements and media choice in collaborative writing. *Human–Computer Interaction*, 7(4):375–407.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Elizabeth Salesky, Julian Mäder, and Severin Klinger. 2021. Assessing evaluation metrics for speech-to-speech translation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 733–740. IEEE.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Matthias Sperber, Ondřej Bojar, Barry Haddow, Dávid Javorský, Xutai Ma, Matteo Negri, Jan Niehues, Peter Polák, Elizabeth Salesky, Katsuhito Sudoh, and Marco Turchi. 2024. Evaluating the IWSLT2023 Speech Translation Tasks: Human Annotations, Automatic Metrics, and Segmentation. ArXiv:2406.03881.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103.

A Head to Head Significance test Results

The following tables (4–6) show differences in average standardized human scores for a system in that column and the system in that row for the German–English language pair. We applied the Wilcoxon rank-sum test to measure the likelihood that such differences could occur simply by chance for text-only, text-audio and WTM22 official[6] experiments. In the following tables, * indicates statistical significance at $p < 0.05$, ** indicates statistical significance at $p < 0.01$, and *** indicates statistical significance at $p < 0.001$, according to the Wilcoxon rank-sum test.

Each table contains a final column showing the total number of judgments used to calculate the results. The number for the official results is much greater than in our experiments; therefore, a direct comparison should only be made between the text-only and multimodal scores.

	HUMAN	JDExploreAcademy	LT22	Lan-Bridge	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT	No. of Judgments
HUMAN	–	-0.19	-0.20	-0.20	-0.31	-0.28	-0.19	-0.38	-0.27	-0.31	371
JDExploreAcademy	0.19**	–	-0.01	-0.01	-0.13	-0.09	0	-0.19	-0.08	-0.12	385
LT22	0.20***	0.01	–	0	-0.12	-0.08	0.01	-0.18	-0.07	-0.11	475
Lan-Bridge	0.20**	0.01	0	–	-0.12	-0.09	0.01	-0.18	-0.07	-0.11	399
Online-A	0.31***	0.13**	0.12*	0.12*	–	0.03	0.13**	-0.07	0.04	0.01	451
Online-B	0.28***	0.09	0.08	0.09	-0.03	–	0.09	-0.10	0.01	-0.02	427
Online-G	0.19**	0	-0.01	-0.01	-0.13	-0.09	–	-0.19	-0.08	-0.12	385
Online-W	0.38***	0.19**	0.18*	0.18**	0.07	0.10*	0.19***	–	0.11*	0.08	433
Online-Y	0.27***	0.08	0.07	0.07	-0.04	-0.01	0.08	-0.11	–	-0.03	417
PROMT	0.31***	0.12*	0.11	0.11	-0.01	0.02	0.12**	-0.08	0.03	–	349

Table 4: Head to Head comparison matrix of text-only judgments with significance levels and number of judgments.

[6]For official results, we used the human evaluation data provided by WMT22 organisers at: https://github.com/wmt-conference/wmt22-news-systems/tree/main/humaneval/DA.

	HUMAN	JDExploreAcademy	LT22	Lan-Bridge	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT	No. of Judgments
HUMAN	–	-0.24	-0.26	-0.21	-0.36	-0.24	-0.39	-0.33	-0.32	-0.35	445
JDExploreAcademy	0.24***	–	-0.02	0.04	-0.12	0	-0.15	-0.09	-0.07	-0.11	426
LT22	0.26***	0.02	–	0.06	-0.10	0.02	-0.13	-0.07	-0.05	-0.09	470
Lan-Bridge	0.21**	-0.04	-0.06	–	-0.15	-0.04	-0.18	-0.13	-0.11	-0.15	514
Online-A	0.36***	0.12*	0.10*	0.15**	–	0.12*	-0.03	0.02	0.04	0	435
Online-B	0.24***	0	-0.02	0.04	-0.12	–	-0.14	-0.09	-0.07	-0.11	499
Online-G	0.39***	0.15*	0.13*	0.18***	0.03	0.14*	–	0.05	0.07	0.03	487
Online-W	0.33***	0.09	0.07	0.13*	-0.02	0.09	-0.05	–	0.02	-0.02	464
Online-Y	0.32***	0.07	0.05	0.11*	-0.04	0.07	-0.07	-0.02	–	-0.04	394
PROMT	0.35***	0.11*	0.09*	0.15**	0	0.11*	-0.03	0.02	0.04	–	549

Table 5: Head-to-head comparison matrix of multimodal annotations with significance levels and number of judgments.

	HUMAN-B	JDExploreAcademy	LT22	Lan-Bridge	Online-A	Online-B	Online-G	Online-W	Online-Y	PROMT	No. of Judgments
HUMAN-B	–	-0.04	0.05***	-0.07	-0.01	0.01	-0.02	-0.05	0.03*	0.07**	2100
JDExploreAcademy	0.04	–	0.09***	-0.03	0.03	0.05*	0.02	-0.01	0.07**	0.11***	2100
LT22	-0.05	-0.09	–	-0.12	-0.06	-0.04	-0.07	-0.10	-0.03	0.02	2100
Lan-Bridge	0.07**	0.03*	0.12***	–	0.06***	0.08***	0.05**	0.02*	0.09***	0.14***	2100
Online-A	0.01	-0.03	0.06**	-0.06	–	0.02	-0.01	-0.04	0.04	0.08**	2100
Online-B	-0.01	-0.05	0.04*	-0.08	-0.02	–	-0.03	-0.06	0.02	0.06*	2100
Online-G	0.02	-0.02	0.07***	-0.05	0.01	0.03	–	-0.03	0.04*	0.09**	2100
Online-W	0.05	0.01	0.10***	-0.02	0.04	0.06*	0.03	–	0.07***	0.12***	2100
Online-Y	-0.03	-0.07	0.03	-0.09	-0.04	-0.02	-0.04	-0.07	–	0.05	2100
PROMT	-0.07	-0.11	-0.02	-0.14	-0.08	-0.06	-0.09	-0.12	-0.05	–	2100

Table 6: Head-to-head comparison matrix of WMT22 official rankings with significance levels and number of judgments.

Meaningful Pose-Based Sign Language Evaluation

Zifan Jiang[1], **Colin Leong***[2], **Amit Moryossef***[1,3],
Oliver Cory[4], **Maksym Ivashechkin**[4], **Neha Tarigopula**[5,6], **Biao Zhang**[7],
Anne Göhring[1], **Annette Rios**[1], **Rico Sennrich**[1], **Sarah Ebling**[1]
[1]University of Zurich [2]University of Dayton [3]sign.mt
[4]University of Surrey [5]Idiap Research Institute [6]EPFL [7]Google DeepMind
jiang@cl.uzh.ch

Abstract

We present a comprehensive study on meaningfully evaluating sign language utterances in the form of human skeletal poses. The study covers keypoint distance-based, embedding-based, and back-translation-based metrics. We show tradeoffs between different metrics in different scenarios through (1) automatic meta-evaluation of sign-level retrieval, and (2) a human correlation study of text-to-pose translation across different sign languages. Our findings, along with the open-source `pose-evaluation` toolkit, provide a practical and reproducible approach for developing and evaluating sign language translation or generation systems.

1 Introduction

Automatic evaluation metrics are essential for assessing the quality of automatically generated language content and tracking progress over time. For instance, machine translation (MT) studies rely heavily on BLEU (Papineni et al., 2002), even though newer metrics have shown stronger correlation with human judgment (Freitag et al., 2022). This trend continues in sign language processing (SLP; Bragg et al. (2019); Yin et al. (2021)), an interdisciplinary subfield of natural language processing and computer vision. Sign language translation (SLT; Müller et al. (2022, 2023a); De Coster et al. (2023)), denoting the part of SLP concerned with translating sign language videos into spoken language text, reuses text-based metrics.

Müller et al. (2023b) puts forward concrete suggestions on evaluating generated text (especially glosses) in a sign language context. They suggest always computing metrics with standardized tools (e.g., SacreBLEU (Post, 2018) for BLEU) and reporting the metric signatures for reproducibility and fair comparison with other work. The opposite direction—generating or translating into sign language utterances (usually from source text)—presents additional challenges for evaluation. Namely, standardized metrics, tooling, and correlation with human evaluation are lacking.

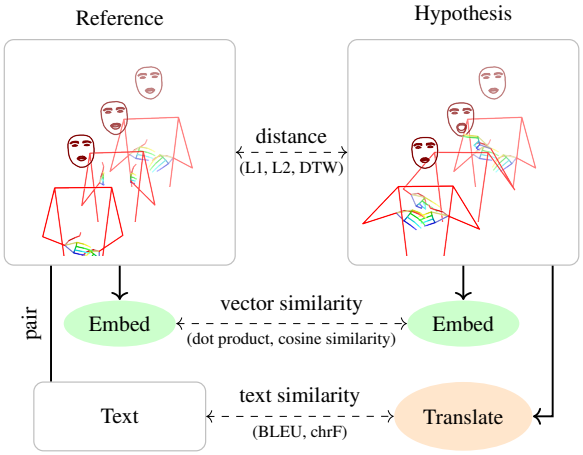


Figure 1: Pose-based evaluation taxonomy overview. We compare a reference and a hypothesis pose sequence by one of the following three ways: (a) computing distance-based metrics directly on the keypoint sequences, optionally aligned by dynamic time wrapping (DTW); (b) encoding each sequence into a shared embedding space and measuring similarity; and (c) back-translating the hypothesis poses into text to apply conventional machine translation metrics on text.

In this work, we systematically examine the metrics employed for evaluating sign language output, especially formatted as human skeletal poses (Zheng et al., 2023) that contain motion of signing (e.g., MediaPipe Holistic; Lugaresi et al. (2019); Grishchenko and Bazarevsky (2020)). We start by a literature review of current research practices in §2, and summarize two major families of metrics: (a) distance-based metrics (§3.1) informed by human motion generation (§2.3) and sign language assessment (SLA; §2.4), assuming the access to reference poses and then computing the distance from the predicted poses to the reference poses, either in the

*Equally contributed as co-second authors.

Proceedings of the Tenth Conference on Machine Translation, pages 64–80
November 8-9, 2025

raw 2D/3D keypoint space or an embedding space; (b) back-translation-based metrics (§3.3) borrowed from MT (Zhuo et al., 2023) and speech translation (Zhang et al., 2023), assuming the pre-existence of a pose-to-text translation model.

After the initial conceptual review, we select, implement, and meta-evaluate typical metrics along with additional innovative ones proposed by us (as summarised in Figure 1), through two empirical approaches: automatic meta-evaluation with a sign-level retrieval task (§4); and a sentence-level correlation study between metrics of interest versus deaf evaluator ratings on three text-to-pose MT systems in three spoken-sign language pairs (§5).

We find that keypoint distance-based metrics, when carefully tuned, can rival more advanced approaches for sign retrieval and human-judgment correlation. On the other hand, embedding-based metrics, including those borrowed from SLA, excel in their own domain but struggle at the sentence level across different systems. Back-translation likelihood emerges as the most consistent metric, highlighting the need for open, standardized pose-to-text models alongside human evaluation.

The source code of the suggested evaluation metrics and the proposed meta-evaluation protocols in §4 are openly maintained in `pose-evaluation`, a public GitHub repository. The human correlation data and evaluation scripts in §5 are also released in a separate `text2pose-human-eval` repository to encourage future research.

2 Related Work

We discuss four related fields in this section with a special emphasis on the evaluation methodology, and outline recent work in sign language generation (SLG; §2.2) in Table 1. The remaining three fields provide additional background relevant to evaluating these SLG systems.

2.1 Sign Language Understanding

Sign language recognition (Adaloglou et al., 2021) and translation (De Coster et al., 2023) are the two most prevalent tasks of understanding sign language from video recordings. The former aims to classify signing into a fixed vocabulary of signs in a particular sign language, either from isolated video clips of single signs (isolated sign language recognition, ISLR) or continuous video footage spanning multiple signs (continuous sign language recognition, CSLR). Given its classification nature, the evaluation efficiently utilizes classic statistical metrics, such as accuracy, F_1 score, and word error rate.

Early SLT attempts rely on glosses (Moryossef et al., 2021b; Müller et al., 2023b), produced manually by humans or a CSLR model. Camgoz et al. (2018, 2020) starts end-to-end neural SLT and leads a wave of gloss-free SLT work (Zhou et al., 2023; Zhang et al., 2024a), where evaluation is typically done with BLEU and BLEURT (Sellam et al., 2020) but not possible with source-based metrics like COMET and quality estimation models like COMET-QE (Chimoto and Bassett, 2022) due to the input modality constraint on sign language. WMT-SLT campaigns for two consecutive years (Müller et al., 2022, 2023a) carry out a rigorous human evaluation process as seen in traditional MT research. Yet the correlation between automatic evaluation metrics and human judgments in SLT has not been reported; quantifying this correlation would yield valuable insights.

2.2 Sign Language Generation

The landscape of SLG is more complicated than SLT, with various inputs, namely, (a) spoken language *text*; (b) sign language *glosses*; (c) iconic *phonetic writing systems* of sign language; (d) textual *phonetic descriptions* of signing, and various outputs, usually, 2D/3D pose; or RGB video frames[1]. We note that in the case of (a) *text*, the generation process involves translation from a spoken language to a sign language with possibly reordering and rephrasing of words, while starting with (b), (c), or (d) merely convert sign language approximated in textual forms into visuals (also known as sign language production[2]), possibly with a preceding step in the pipeline that translates from (a) *text* to (b) *SignWriting* (Jiang et al., 2023), (c) *glosses* (Zhu et al., 2023), or (d) *descriptions*[3]. Our work evaluates poses as the primary representation of sign language motion and semantics, deliberately excluding RGB videos to avoid confounding factors such as visual appeal or signer identity. Evaluating videos using the same methods is possible after first estimating them into poses.

[1]For further details about these representations, please refer to the explanatory figures on https://research.sign.mt/.

[2]The terms are sometimes used interchangeably and thus confuse. This work adheres to the broad term of sign language generation, which involves generating signing from any source.

[3]For example, signing HELLO in ASL: dominant B-hand at forehead → short outward stroke; friendly/smiling face.

Work	Datasets	Sources	Target	Model		Evaluation Metrics					
	(P,H2, etc.)	(T,G,H)	(M,O,S)	</>	θ	D	</>	B4	</>	θ	Other
Arkushin et al. (2023)	DGS Corpus, 3 others	H	O	✔	✔	✔	✔	n/a	n/a	n/a	-
Stoll et al. (2018, 2020)	P	G	O	-	n/a	-	-	-	-	-	SSIM, PSNR, MSE (pixel-wise)
Moryossef et al. (2023b)	Signsuisse	G	M	✔	n/a	-	-	-	-	-	-
Zuo et al. (2024b)	P,CSL-Daily	G	S	✔	n/a	✔	-	✔	✔	✔	Frame temporal consistency
Saunders et al. (2020b,a, 2021a,b)	P	T	O	✔	-	-	-	✔	✔	-	-
Hwang et al. (2021, 2023)	P,H2	T	O	✔	-	✔	-	✔	-	-	Fréchet Gesture Distance
Yin et al. (2024)	P	T	S	-	-	✔	-	✔	-	-	-
Fang et al. (2024a,b)	P,H2,4 others	T	O	-	-	✔	-	✔	-	-	SSIM, Hand SSIM, FID, etc.
Yu et al. (2024)	P,H2,4 others	T,G,H	S	✔	-	✔	-	-	-	-	FID, Diversity, MM-Dist, etc.
Baltatzis et al. (2024)	H2	T	S	-	-	✔	-	✔	-	-	FID
Zuo et al. (2024a)	P,H2,CSL-Daily	T	S	-	-	✔	-	✔	-	-	Latency

Table 1: Literature review of recent works on pose-based sign language generation (May 2025). P=RWTH-PHOENIX-Weather2014T, H2=How2Sign; T=Text, G=Gloss, H=HamNoSys; M=MediaPipe, O=OpenPose, S=SMPL-X; D=DTW-MJE (and other distance-based metrics), B4=BLEU-4 (and other back translation-based metrics); </> and θ represent the availability of source code and model weights for the generation model and the evaluation metrics (including the back translation model if involved), respectively. The check mark symbols (✔) are clickable links in these columns, and *n/a* denotes not-applicable cases, such as model weights for gloss-based systems and back-translation for HamNoSys input. Other image-based metrics are left as less relevant.

We present prominent pose-based SLG studies from recent years, along with their evaluation methods, in Table 1, grouped by input modalities. Following a similar roadmap as SLT, SLG takes off with a gloss-based cascading approach (text-to-gloss-to-sign; Stoll et al. (2018, 2020)) and then gradually switches to an end-to-end fashion in a series of follow-up work (Saunders et al., 2020b,a, 2021a,b). Attempts have also been made with alternative phonetic inputs such as HamNoSys (Prillwitz and Zienert, 1990; Arkushin et al., 2023).

Unlike SLT, however, gloss-based baseline approaches for SLG remain competitive and practical choices (Moryossef et al., 2023b; Zuo et al., 2024b) due to accessible sign language dictionary resources that enable straightforward mapping of glosses to sign language pose sequences. Modern end-to-end approaches utilise vector quantization, diffusion models, and LLMs, and the output pose format spans from classic 2D standards such as MediaPipe Holistic and Openpose (Cao et al., 2019) to 3D SMPL-X (Pavlakos et al., 2019).

Popular datasets used in this line of work include RWTH-PHOENIX-Weather 2014T, in German Sign Language (DGS), introduced by Forster et al. (2014); Camgoz et al. (2018); CSL-Daily, in Chinese Sign Language (CSL), introduced by Zhou et al. (2021); and How2Sign, in American Sign Language (ASL), introduced by (Duarte et al., 2021). We choose Signsuisse (Müller et al., 2023a) in this work (§5) for its multilingual nature and richer vocabulary than others[4].

As for evaluation, the SLRTP Sign Language Production Challenge 2025 summarises the most common evaluation metrics: (a) keypoint distance-based, such as DTW-MJE (Dynamic Time Warping - Mean Joint Error); and (b) back-translation-based, such as BLEU and BLEURT. Human evaluation is conducted briefly in Saunders et al. (2021a,b), Baltatzis et al. (2024), and Zuo et al. (2024a) and more extensively in another concluded campaign–Quality Evaluation of Sign Language Avatars Translation (Yuan et al., 2024). Unfortunately, like in SLT, the correlation between automatic metrics and human judgments has never been formally validated. Upon reviewing Table 1, we spot two significant issues in the current development of SLG: (a) Most systems and their evaluations are non-reproducible due to the lack of source code and model weights (including the back-translation models if involved). (b) Cross-work comparisons are unrealistic given the fragmented implementation of the evaluation metrics (in contrast to MT, where standardized tools like SacreBLEU are available).

2.3 Human Motion Generation

Motion generation from natural language is a related field where human pose sequences are synthesized to reflect described actions (Tevet et al., 2022; Zhang et al., 2024b). Evaluation typically involves distance-based metrics (e.g., joint or velocity error), perceptual similarity (e.g., Fréchet Inception Distance adapted to motion), and alignment metrics, such as R-Precision, to measure text-motion coherence. However, Voas et al. (2023) shows that many of these automated metrics correlate poorly

[4]PHOENIX and CSL-Daily feature 1066 and 2000 signs.

with human judgment on a per-sample basis. They propose MoBERT, a BERT-based learned evaluator, which achieves higher agreement with human ratings, highlighting the ongoing challenge of designing semantically meaningful motion evaluation.

2.4 Sign Language Assessment

SLA research compares student-produced signing against canonical references. Cory et al. (2024) evaluates sign language proficiency by modeling the natural distribution of signing motion across multiple references and demonstrating a strong correlation with human ratings. Tarigopula et al. (2024, 2025) proposes a posterior-based analysis of skeletal or spatio-temporal features to assess both manual and non-manual signing components, improving alignment with human evaluation as well.

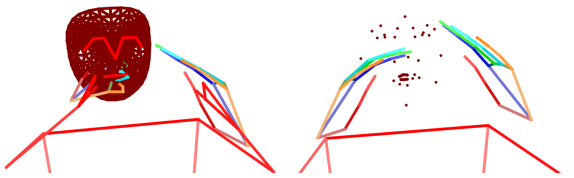
3 Evaluation Metrics

In this section, we formally define common evaluation metrics mentioned in related work (§2) and implement them, reusing open-source code where available, to prepare for the upcoming empirical study on pose evaluation in §4 and §5.

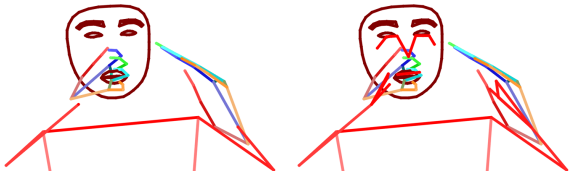
3.1 Keypoint Distance-Based Metrics

We borrow keypoint distance-based metrics from prior work on sign language generation, notably Ham2Pose (Arkushin et al., 2023). These metrics, e.g., APE (Average Position Error)—initially developed for general pose estimation and motion analysis—quantify geometric similarity using frame-wise errors and alignment strategies. However, they are not designed for sign language and ignore critical linguistic properties such as signer speed variation, hand dominance, and missing keypoints. Moreover, they have not been systematically validated against human judgments in sign language contexts, motivating our extended investigation.

During (re-)implementation, we identify significant sources of variation that affect the outcomes of distance-based metrics: (a) whether and how the coordinate values of the keypoints are normalized (e.g., based on the shoulder position as the origin $(0, 0)$ and the shoulder width being 1); (b) whether videos are trimmed to exclude signing-inactive frames; (c) which subset of the keypoints from the pose estimation library is selected for comparison (Figure 2; e.g., hands-only vs. full body); (d) how framerate mismatches are handled (e.g., interpolating to a consistent FPS); (e) how masked

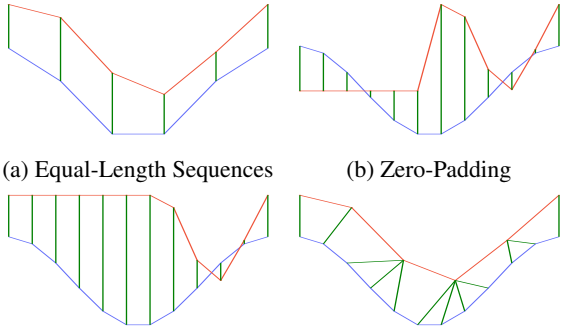


(a) Full (586 Points) (Grishchenko and Bazarevsky, 2020)
(b) YouTube-ASL (85 Points) (Uthus et al., 2023)



(c) Face-contour (178 Points) (Moryossef et al., 2021a)
(d) SignCLIP (203 Points) (Jiang et al., 2024)

Figure 2: MediaPipe keypoint selection strategies.



(a) Equal-Length Sequences
(b) Zero-Padding
(c) Pad with First Frame
(d) Dynamic Time Warping

Figure 3: Sequence alignment (in green) between a shorter sequence (in red) and longer sequence (in blue). In reality, pose keypoint trajectories are aligned temporally in 3D and then averaged for the whole body. Paddings take values from the first frame or simply 0s.

or missing keypoints are treated (e.g., filled with a value, or a default distance returned, or simply ignored); and (f) how sequences of unequal length are aligned before applying APE (Figure 3; e.g., using zero-padding, frame repetition, or DTW).

We examine these variations and provide a reproducible toolkit that enables tuning these design choices explicitly—including keypoint selection, normalization, and masking, sequence trimming and alignment with different distance measures—rather than inheriting arbitrary defaults. The toolkit supports the generation of possibly thousands of metric variants to be tested in §4.

3.2 Embedding-Based Metrics

Rather than operating on the keypoints' raw spatial positions, we categorize embedding-based metrics that calculate distance or similarity in a latent em-

bedding space provided by a trained model.

3.2.1 Sign Language Assessment Metrics

We adopt two metrics for comparing two poses from the SLA task (§2.4): the Skeleton Variational Autoencoder (SkeletonVAE) model from Cory et al. (2024) and the posterior-based scores from assessment models developed in Tarigopula et al. (2024).

SkeletonVAE Score The SkeletonVAE is trained to produce a per-frame latent embedding. 2D MediaPipe poses are first uplifted to constrained 3D skeletons using the method of Ivashechkin et al. (2023) and then embedded into a 10-dimensional β-VAE latent space (Higgins et al., 2017). We define *SkeletonVAE Score* as the L2 distance between the reference and hypothesis sequences' DTW-aligned latent trajectories, optionally normalized by the DTW path length.

Skeleton Posterior-based SKL Score Following Tarigopula et al. (2024), we first extract two sets of linguistically informed features from the pose sequences with the same missing keypoint preprocessing as Eq. 6 in Arkushin et al. (2023). For hand movement, we compute 36-dimensional feature vectors representing hand position and velocity relative to the head, shoulders, and hips with a temporal context of 9 frames. For handshape, we calculate joint positions relative to the wrist and input them into a separate MLP to obtain handshape posteriors. The resulting stack of shape and movement posteriors from both the reference and hypothesis examples is then aligned using DTW with a cost function based on the Symmetric Kullback–Leibler (SKL) divergence. The cost is aggregated over the DTW time steps as the final score with two variants–*SKL_mvt Score* (movement only) and *SKL_mvt_hshp Score* (movement + handshape), respectively.

3.2.2 SignCLIP Score

One step further than §3.2.1, we follow *CLIPScore* (Hessel et al., 2021) and use SignCLIP (Jiang et al., 2024), a model repurposed for representing sign language poses by multilingual contrastive learning, to derive *SignCLIPScore P-P* (pose-to-pose), based on the dot product of the embeddings of the reference and hypothesis on the example level instead of frame-level latents plus DTW alignment.

Reference-Free Quality Estimation Variant We introduce *SignCLIPScore P-T* (pose-to-text). It computes the dot product between the text and pose embedding, eliminating reliance on scarce or even unreliable ground-truth signing references (Freitag et al., 2023).

3.3 Back-Translation-Based Metrics

Assuming the existence of the corresponding spoken language text and a reliable pose-to-text SLT model, we can evaluate a sign language pose by: (a) *Sampling*: translate the pose sequence into text, then compare with the source text using BLEU[5], chrF[6], or BLEURT. (b) *Scoring*: compute the log-likelihood of the text given the pose sequence as input to the SLT model. This avoids errors introduced by decoding and supports more consistent comparisons across systems. In this study, we adopt an SLT model from Zhang et al. (2024a), which is pretrained on a large-scale YouTube SLT corpus and massive MT data. We use *system 8* from their study (*YT-Full + Aug-YT-ASL&MT-Large + ByT5 XL*), i.e., the current state of the art, and preprocess the generated pose sequences by selecting the same 85 keypoints specified in their paper[7].

4 Automatic Meta-Evaluation

In this section, we explore methods to automatically (meta-)evaluate proposed metrics, especially when there are many variants as seen in §3.1.

We adopt the retrieval-based evaluation protocol from Arkushin et al. (2023) to assess how well different metrics capture meaningful distinctions between signs. Each pose sequence is treated as a query, and the goal is to retrieve other samples of the same sign (*targets*) from a pool that includes unrelated signs (*distractors*). We focus primarily on the distance-based metric variants introduced in §3.1, and compare them against embedding-based alternatives such as *SignCLIP Score* (§3.2.2).

Evaluation is conducted on a combined ASL dataset of ASL Citizen (Desai et al., 2023), SemLex (Kezar et al., 2023), and PopSign ASL (Starner et al., 2023). For each sign/gloss, we use all available samples as targets and sample four times as many distractors, yielding a 1:4 target-to-distractor ratio. For instance, for the sign *HOUSE* with 40 samples (11 from ASL Citizen, 29 from Sem-Lex), we add 160 distractors and compute pairwise dis-

[5] `nrefs:1|case:mixed|eff:yes|tok:13a|smooth:exp|version:2.3.1`

[6] `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`

[7] Eight mismatched keypoints due to different MediaPipe versions are imputed as missing landmarks.

Base (f)	Fill (e)	Trim (b)	Norm. (a)	Padding (f)	Keypoints (c)	**mAP↑**	**P@10↑**
Ham2Pose nAPE	0*	✗	✔	zero	Reduced	26%	14%
Ham2Pose nDTW(-MJE)	unspecified	✗	✔	/	Reduced	27%	14%
APE	10	✗	✗	zero	Upper body	33%	27%
APE	10	✗	✗	first-frame	Upper body	34%	29%
APE	10	✔	✗	zero	Upper body	35%	30%
APE	10	✗	✔	zero	Reduced	36%	32%
APE	10	✗	✗	first-frame	Reduced	37%	32%
APE	10	✗	✗	first-frame	YT-ASL	39%	36%
APE	10	✔	✗	first-frame	Reduced	40%	36%
APE	10	✔	✔	zero	Upper body	41%	37%
APE	10	✗	✗	zero	Hands	42%	38%
APE	10	✔	✔	first-frame	YT-ASL	43%	39%
APE	10	✔	✗	zero	Hands	45%	41%
DTW	10	✗	✗	/	Upper body	36%	32%
DTW	10	✔	✗	/	Upper body	37%	33%
DTW	10	✗	✗	/	Reduced	42%	40%
DTW	10	✗	✔	/	Upper body	43%	41%
DTW	10	✔	✔	/	Upper body	43%	41%
DTW	10	✗	✔	/	Reduced	44%	41%
DTW	10	✗	✔	/	Hands	45%	41%
DTW	10	✗	✗	/	YT-ASL	48%	47%
DTW	10	✔	✗	/	YT-ASL	49%	48%
DTW	10	✗	✗	/	Hands	53%	52%
DTW‡	10	✔	✗	/	Hands	53%	52%
DTW†	1	✗	✔	/	Hands	55%	53%
SignCLIPScore P-P (multilingual)						50%	48%
SignCLIPScore P-P (ASL finetuned)						**91%**	**92%**

Table 2: Automatic meta-evaluation of reference-based evaluation metrics on sign retrieval across various settings: (a)-(f) enumerated in section 3.1. The table presents a representative subset of top-performing metrics. *Fill* indicates the value used to fill in missing keypoint; *zero* indicates zero-padding; *first-frame* indicates padding with the first frame. *YT-ASL* includes a subset of keypoints used and described in Uthus et al. (2023). *Reduced* includes a subset of keypoints used and described in Jiang et al. (2024). * Ham2Pose nAPE implements missing-filling slightly differently–filling in zeros for both trajectories if one of them has a missing value (see details in Appendix A).

tance from each target to all 199 other examples. The pairwise distance is defined by each of these proposed metric scores. Retrieval quality is measured using Mean Average Precision (mAP↑) and Precision@10 (P@10↑). The full evaluation covers 5362 unique signs and 82,099 pose sequences. After several pilot runs to rule out clear bad choices, we finalize a subset of 169 signs with at most 20 samples each, and evaluate 48 representative keypoint distance-based metric candidates and *SignCLIP Score* with different SignCLIP checkpoints provided by the authors[8] on this subset. For reference, we also reproduce the metrics proposed by Arkushin et al. (2023). The key results, including the best metrics, are presented in Table 2.

[8] https://github.com/J22Melody/fairseq/tree/main/examples/MMPT#demo-and-model-weights

The results show that, as expected, DTW-based metrics outperform padding-based APE baselines. While selecting hands-only keypoints appears to yield the best results, a more sophisticated selection that includes non-manuals might still be desirable. Embedding-based methods, particularly SignCLIP models fine-tuned on in-domain ASL data, achieve the strongest retrieval scores. We mark the two best DTW-based metrics by ‡ and †, and rename them *DTWp* and *nDTWp* for use in the rest of the paper.

5 Text-to-Pose Translation Study with Human Evaluation

This section shifts our evaluation focus from automatic sign-level tasks to a sentence-level text-to-pose sign language machine translation scenario. Due to their subjective and diverse nature, open-

ended text or utterance generation tasks inherently lack a single "correct"/"ground-truth" answer. Consequently, **automatic evaluation metrics are only *meaningful* if they correlate closely with human judgments** (Reiter, 2018; Sellam et al., 2020).

5.1 Dataset: WMT-SLT Signsuisse

We use the Signsuisse dataset released in the WMT-SLT 23 campaign. The dataset comprises 18,221 lexical items in three spoken-sign language pairs, represented as videos and glosses. One signed example sentence for each lexical item is presented in a video along with the corresponding spoken language translation, which forms parallel data between the sign and spoken languages. The test set is used to test different text-to-pose translation systems. It contains 500 German/Swiss German Sign Language (DSGS) segments, 250 French/French Sign Language (LSF) segments, and 250 Italian/Italian Sign Language (LIS) segments.

5.2 Systems

We utilize three text-to-pose translation systems that convert spoken language text inputs into corresponding sign language represented by the MediaPipe Holistic pose formats.

Reference* MediaPipe poses are estimated from the reference translation videos, i.e., ground truth.

sign.mt Based on Moryossef et al. (2023b), this open system converts text into sign language glosses through rule-based reordering and selective word dropping. Glosses are mapped to skeletal poses retrieved from a lexicon and are then concatenated to form coherent sequences. When a gloss is missing from the lexicon, the system defaults to fingerspelling the corresponding word.

sign.mt v2 During evaluation, we found that frequent fingerspelling of missing glosses was cumbersome and frustrating for evaluators. Therefore, in this version, we opted to omit glosses without lexical mappings, acknowledging that while this may result in information loss, it significantly improves user experience and evaluation efficiency.

Sockeye We adapt Sockeye (Hieber et al., 2022) to continuous pose sequences by modifying both the encoder and decoder to handle continuous sequences. The text-to-pose Sockeye model is trained on the Signsuisse training set with 60k updates on a 32GB NVIDIA Tesla V100 GPU.

To avoid exposure bias—where the decoder overfits to gold frames and fails at inference, we first predict only the initial pose y_1 from the encoder output, then feed y_1 as input for all subsequent steps $y_{2:n}$, training the decoder to output frame-to-frame deltas $\Delta y_t = y_t - y_1$ instead of absolute poses. Since the target sequence is continuous, we replace the cross-entropy loss function with mean squared error on the poses. Additionally, there is no `<EOS>` token with continuous output; instead, we learn to output the length of the pose sequence based on the length ratios from the training data. We provide the link to the adapted Sockeye repository and a demo of translation output.

5.3 Human Evaluation

We collect system translations and use Appraise (Federmann (2018); Figure 4) to allow evaluators to rate the translations on a continuous scale between 0 and 100 as in traditional direct assessment (Graham et al., 2013; Cettolo et al., 2017) but with 0-6 markings on the analogue slider and custom annotator guidelines designed explicitly for our task (similar to WMT-SLT, but reverse translation direction). Evaluation instructions are sent out in DSGS, LSF, and LIS, which are translations of the respective spoken language instructions in WMT-SLT. The instructions are attached in Appendix B.

We hire seven DSGS, two LSF, and four LIS evaluators, all of whom are native deaf sign language users[9]. All work is done with informed consent in written and signed form. Of the seven native DSGS deaf signers, four have never participated in such an evaluation campaign before, two have participated once, and one has attended more than once. Concerning their professional backgrounds, four are deaf translators; one also interprets live. Complete demographics are presented in Table 4.

An initial round of evaluation informs us about the cost, roughly 100 example segments per hour, with a compensation of ∼40 USD per hour. Evaluators also provide constructive feedback on the Appraise platform and the translation systems, which results in the switching into the v2 version of sign.mt. Therefore, the number of evaluated examples varies slightly between systems and languages.

Statistics The evaluation comprises 2650 unique examples and 11,471 ratings across all four systems (3275 reference, 4032 Sockeye, 861 sign.mt,

[9]One additional DSGS evaluator, a hearing interpreter, did a pilot study with us to test the Appraise system.

	Reference-Based								Reference-Free						
	Distance-Based				SLA Metrics				SignCLIPScore		Back Translation-Based				H*
	nAPE	**nDTW**	**DTWp**	**nDTWp**	**SVAE**	**SVAE$_n$**	**SKL**	**SKL$_h$**	**P-P**	**P-T**	**B4**	**chrF**	**B-RT**	**Lik.**	**H***
By System															
sign.mt	0.09	0.14	0.11	0.10	0.23	-0.08	0.24	0.17	0.10	0.02	0.05	0.11	0.05	0.23	0.43
sign.mt v2	0.28	0.33	0.26	0.31	0.46	0.14	0.22	0.29	0.00	-0.19	0.20	0.22	0.44	0.49	0.52
Sockeye	0.10	0.15	0.04	0.17	0.13	0.01	0.24	0.07	0.42	-0.27	-0.07	0.04	0.46	0.58	0.22
By Language															
DE→DSGS	-0.36	-0.09	0.73	0.43	-0.02	0.27	-0.57	-0.51	-0.31	0.39	0.18	0.26	0.09	0.36	0.70
FR→LSF	-0.54	-0.11	0.76	0.02	-0.01	0.37	-0.68	-0.65	-0.01	0.45	0.32	0.60	0.47	0.29	0.80
IT→LIS	-0.57	-0.39	0.79	0.57	-0.02	0.53	-0.75	-0.74	0.13	0.29	0.31	0.63	0.41	0.38	0.88
Overall (↑)	-0.41	-0.10	0.76	0.43	0.07	0.38	-0.56	-0.53	-0.10	0.27	0.21	0.42	0.36	0.42	0.77
SD (↓)	(0.35)	(0.24)	(0.34)	(0.20)	(0.18)	(0.22)	(0.47)	(0.43)	(0.22)	(0.29)	(0.14)	(0.23)	(0.18)	(0.12)	(0.24)

Table 3: Segment-level Spearman correlations with average human judgments calculated for several pose-based evaluation metrics for sign language. nAPE=`normalized APE`, nDTW=`normalized DTW-MJE` (two metrics taken from Arkushin et al. (2023) and re-implemented for MediaPipe, normalized by pose shoulder); DTWp=`DTW+Trim+MaskFill10.0+Hands-Only`, nDTWp=`DTW+MaskFill1.0+Norm.+Hands-Only` (top metrics selected in §4 implemented by `pose-evaluation`, denoted by ‡ and † in Table 2, without/with pose normalization, respectively); SVAE=`SkeletonVAE Score`, SVAE$_n$=`SVAE normalized by DTW path`, SKL=`SKL_mvt Score`, SKL$_h$=`SKL_mvt_hshp Score`; P-P=`Pose-to-pose embedding distance`, P-T=`Pose-to-text embedding distance`; B4=`BLEU-4`, chrF=`chrF`, B-RT=`BLEURT`, Lik. =`Likelihood`. H* denotes mean inter-evaluator Spearman correlation. SD represents the standard deviation across each column and is expected to be small for an ideal metric.

and 3303 sign.mt v2) and three language pairs (7861 DSGS, 1210 LSF, and 2400 LIS).

We follow the practices set by WMT-SLT. The inter-annotator agreement, measured with an approximation of Fleiss κ (Fleiss, 1971) by discretizing the continuous scale 0-100 in seven bins in the scale 0-6, is $\kappa = 0.36 \pm 0.05$. We also randomly mix 500 references and some repeated hypothesis segments for sanity checks and quality control. The mean intra-annotator agreement over all evaluators is $\kappa = 0.49 \pm 0.09$, calculated over 50-100 segments evaluated twice by the same evaluator. We find the inter- and intra-annotator agreement to be lower than in the WMT-SLT study for the sign-to-text translation direction, and posit that the lack of a clear definition and criteria for translation quality in sign language poses a significant challenge.

	Evaluation experience			SL professional			Avg yr.
	Never	Once	> Once	Translator	Interpreter	Teacher	
DSGS (7)	4	2	1	4	1	4	39.0
LSF (2)	0	1	1	1	0	1	35.0
LIS (4)	1	1	2	3	4	0	42.5

Table 4: Raters overview: system evaluation and professional experience with sign language, average number of years signing (in most cases equivalent to age).

5.4 Correlation Analysis

We perform a correlation analysis between the metrics proposed in §3 and the human scores averaged over evaluators at the segment level, as presented in Table 3. The metrics are divided into families, where *reference-based* means that the quality of the translated poses is measured in relation to reference poses derived from signing videos. The absolute scores per metric/system are presented in Table 5.

For the distance-based metrics, we reproduce *nAPE* and *nDTW(-MJE)* for MediaPipe poses based on the open implementation from Arkushin et al. (2023) as a reference, and additionally compare them to the best-performing metrics informed by the automatic meta-evaluation in §4 on ASL, a different sign language. We flip the signs of the metrics that quantify errors to keep a positive correlation for analytical convenience. Row-wise, we first break down the correlation by system and language into relevant rows, and then present the overall correlation, including all systems and languages, to reflect performance at the system level.

6 Discussion and Recommendations

Distance-based metrics are efficient defaults, but the devil is in the implementation details. Although seemingly straightforward to implement, distance-based metrics involve many design choices, including pose format and keypoint selection. We empirically demonstrate the effectiveness of correcting these choices through a random parameter search, following our established meta-evaluation protocols in §4. We recommend using the tuned versions—*DTWp* and

	nAPE↓	nDTW↓	DTWp↓	nDTWp↓	SVAE↓	SVAE$_n$↓	SKL↓	SKL$_h$↓	P-P↑	P-T↑	B4↑	chrF↑	B-RT↑	Lik.↑	H*
reference*	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	74.23	15.05	38.52	0.49	-32.87	76.55
sign.mt	0.60	25.43	5171.55	8.66	1.20	0.0028	2794.29	7443.66	81.58	75.55	3.46	14.53	0.26	-39.30	22.00
sign.mt v2	1.65	20.73	4508.62	8.97	1.23	0.0045	4165.78	11540.15	89.89	76.46	6.89	24.79	0.23	-67.55	30.12
Sockeye	0.29	16.97	11879.58	10.71	1.16	0.0057	1056.29	3985.65	94.45	72.40	3.44	11.90	0.17	-77.57	5.05

Table 5: Mean absolute scores for each metric across systems. Rows and columns mirror those in Table 3.

nDTWp—in our `pose-evaluation` library, or tuning your distance-based metrics when necessary.

Upon successful tuning, a distance-based metric achieves decent sign retrieval and correlation with humans in text-to-pose translation. Our tuned metrics can even be used as a distance function for a nearest neighbor classifier[10], and reach close performance as the SignCLIP model pretrained on multilingual sign language data; still, it lags behind a SignCLIP model fine-tuned on in-domain data (Table 2). When used to evaluate translation output, keypoint distance-based metrics can range from negatively correlated with human judgments (as seen for *nAPE* and *nDTW*), to being the best metrics tested. *DTWp* wins the overall correlation while *nDTWp* is more sensitive on the segment level within a specific system (Table 3).

SLA metrics correlate with humans on the segment level but are confused on the system level. While verified to align with human ratings for their tasks on evaluating human-produced signing (usually involving fixed individual signs), text-to-pose translation is more lengthy and open-ended, which hinders direct transferability. A proper length normalization (as seen in the case of *SVAE*$_n$ vs. *SVAE*) might help on the system level at the price of losing precision on the segment level.

SignCLIP, used as a multilingual embedding device, excels on the sign level, but falls short for sentence-level translation evaluation. We speculate that using a single embedding to summarize a long-duration ($>$ 10 seconds) signing video is inherently limited, especially for DSGS, a language unseen during SignCLIP's pretraining. Nevertheless, the reference-free variant exhibits a moderate correlation at the system level, and we observe a similar tradeoff (*P-P* vs. *P-T*) between segment and system level correlations, as seen in the SLA metrics.

[10]Upon quick experimentation, *nDTWp* with KNN (n=10) achieves 19% ISLR accuracy on the ASL Citizen test set. We leave a more systematic evaluation on this end to future work.

Back-translation-based approaches correlate properly with human judgment; a gap remains compared to inter-human correlation. In addition to the standard practices suggested by Müller et al. (2023b) on computing text-based metrics, we call for open, standardized pose-to-text translation models that include both the model weights and the source code. Yet, as noted in Table 1, this is hardly the case in current research, and having a dedicated back-translation model for each translation direction (or even dataset) is a luxury. The above-mentioned metrics, which do not rely on in-domain data but function to a decent degree, are valuable in a more generic setting. Human evaluation shall be used as the final quality assessment resort.

When using back translation, likelihood is consistent and more reliable than text metrics. BLEU, chrF, and BLEURT show weaker or unstable correlations with humans in Table 3. It is recommended that back-translation likelihood be included as a primary metric when a pose-to-text model is available.

7 Conclusion

This work presents a unified framework and an open-source `pose-evaluation` toolkit for systematically assessing (generated) sign language utterances based on human skeletal poses. We implemented and compared a wide range of metrics (§3)—distance-based, embedding-based, and back-translation-based—via automatic meta-evaluation on sign retrieval (§4) and a comprehensive human correlation study across three sign languages (§5). Our results demonstrate that carefully tuned distance metrics, namely *DTWp* and *nDTWp*, and back-translation likelihoods yield the strongest agreement with native signer judgments. We release our code, evaluation protocols, and human ratings to foster reproducible and fair comparisons in computational sign language research.

8 Limitations

8.1 3D Pose Representation

While our study focuses on using MediaPipe Holistic as the pose format for representing sign language motion, other specifics, especially the recently developed 3D SMPL-X (Pavlakos et al., 2019) would be a visually more expressive choice. However, the lack of a common way to extract and use 3D poses as easily as MediaPipe Holistic makes the latter the most used choice in SLP.

8.2 Missing Publicly Available Systems

Our study is further limited by the number of public systems (Table 1) we can use to run the correlation analysis, unless we implement everything from scratch (including the pose estimation pipelines, text-to-pose systems, and back-translation models). We hope the release of this work will alleviate the situation.

8.3 Automatic Evaluation beyond Sign Level

The automatic meta-evaluation in §4 is capped by the sign-level retrieval task, and we envision extending it to the phrase level. One possible approach is to leverage the Platonic Representation Hypothesis proposed by Huh et al. (2024). In the pose evaluation scenario, we hypothesize that the similarity given by a good pose metric between two pose segments should correlate with the similarity given by a text embedding model between the two text segments paired with the two pose segments, respectively. We leave exploration on this end for future work, which will likely connect the automatic meta-evaluation more closely to the sentence-level human correlation study in §5.

8.4 Tokenized Evaluation

Inspired by how text metrics like BLEU collect surface-form overlapping statistics, we envision a tokenized evaluation as promising for sign language evaluation. Although a sign language pose sequence cannot be discretely tokenized and matched like text tokens, the combination of a sign language segmentation model (Moryossef et al., 2023a) plus SignCLIP embedding can be utilized in a way similar to BERTScore (Zhang* et al., 2020), where a similarity matrix is constructed between the reference and hypothesis tokens to derive the final similarity score on phrase level.

Acknowledgments

This work is funded by the Swiss Innovation Agency (Innosuisse) flagship IICT (PFFS-21-47) and by the SIGMA project at the UZH Digital Society Initiative (DSI).

We thank the deaf evaluators for their efforts in the human evaluation and valuable feedback on the evaluated systems. We thank SWISS TXT for helping organize the evaluation campaign. We thank Roman Grundkiewicz for troubleshooting the Appraise platform and thank Lisa Arter for a pilot test on it. We thank Andreas Säuberli for the insightful discussion on inter-annotator agreement. We also thank Garrett Tanzer for answering questions about the MediaPipe version used in Google's pose-to-text translation system, and Ronglai Zuo for the advice on the paper draft.

References

Nikolas Adaloglou, Theocharis Chatzis, Ilias Papastratis, Andreas Stergioulas, Georgios Th Papadopoulos, Vassia Zacharopoulou, George J Xydopoulos, Klimnis Atzakas, Dimitris Papazachariou, and Petros Daras. 2021. A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, 24:1750–1762.

Rotem Shalev Arkushin, Amit Moryossef, and Ohad Fried. 2023. Ham2pose: Animating sign language notation into pose sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21046–21056.

Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. 2024. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, et al. 2019. Sign language recognition, generation, and translation: An interdisciplinary perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 16–31.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition

and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Everlyn Asiko Chimoto and Bruce A. Bassett. 2022. COMET-QE and active learning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4735–4740, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Oliver Cory, Ozge Mercanoglu Sincan, Matthew Vowels, Alessia Battisti, Franz Holzknecht, Katja Tissi, Sandra Sidler-Miserez, Tobias Haug, Sarah Ebling, and Richard Bowden. 2024. Modelling the distribution of human motion for sign language assessment. In *Proceedings of the 12th Workshop on Assistive Computer Vision and Robotics (ACVR) at ECCV*.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2023. Machine translation from signed to spoken languages: State of the art and challenges. *Universal Access in the Information Society*, pages 1–27.

Aashaka Desai, Lauren Berger, Fyodor O. Minakov, Vanessa Milan, Chinmay Singh, Kriston Pumphrey, Richard E. Ladner, Hal Daum'e, Alex X. Lu, Naomi K. Caselli, and Danielle Bragg. 2023. Asl citizen: A community-sourced dataset for advancing isolated sign language recognition. *ArXiv*, abs/2304.05934.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sen Fang, Chunyu Sui, Yanghao Zhou, Xuedong Zhang, Hongbin Zhong, Minyu Zhao, Yapeng Tian, and Chen Chen. 2024a. Signdiff: Diffusion models for american sign language production. *Preprint*, arXiv:2308.16082.

Sen Fang, Lei Wang, Ce Zheng, Yapeng Tian, and Chen Chen. 2024b. Signllm: Sign languages production large language models. *Preprint*, arXiv:2405.10718.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Ivan Grishchenko and Valentin Bazarevsky. 2020. Mediapipe holistic.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, et al. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv preprint arXiv:2207.05851*.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR.

Eui Jun Hwang, Jung-Ho Kim, and Jong C Park. 2021. Non-autoregressive sign language production with gaussian space. In *BMVC*, volume 1, page 3.

Eui Jun Hwang, Huije Lee, and Jong C Park. 2023. Autoregressive sign language production: A gloss-free approach with discrete representations. *arXiv preprint arXiv:2309.12179*.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. 2023. Improving 3d pose estimation for sign language. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023. Machine translation between spoken languages and signed languages represented in SignWriting. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1706–1724, Dubrovnik, Croatia. Association for Computational Linguistics.

Zifan Jiang, Gerard Sant, Amit Moryossef, Mathias Müller, Rico Sennrich, and Sarah Ebling. 2024. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, Miami, Florida, USA. Association for Computational Linguistics.

Lee Kezar, Elana Pontecorvo, Adele Daniels, Connor Baer, Ruth Ferster, Lauren Berger, Jesse Thomason, Zed Sevcikova Sehyr, and Naomi Caselli. 2023. The sem-lex benchmark: Modeling asl signs and their phonemes. *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

Amit Moryossef, Zifan Jiang, Mathias Müller, Sarah Ebling, and Yoav Goldberg. 2023a. Linguistically motivated sign language segmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12703–12724, Singapore. Association for Computational Linguistics.

Amit Moryossef, Mathias Müller, and Rebecka Fahrni. 2021a. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/sign-language-processing/pose`.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023b. An open-source gloss-based baseline for spoken to signed language translation. In *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*, pages 22–33, Tampere, Finland. European Association for Machine Translation.

Amit Moryossef, Kayo Yin, Graham Neubig, and Yoav Goldberg. 2021b. Data augmentation for sign language gloss translation. In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 1–11, Virtual. Association for Machine Translation in the Americas.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Siegmund Prillwitz and Heiko Zienert. 1990. Hamburg notation system for sign language: Development of a sign writing with computer application. In *Current trends in European Sign Language Research. Proceedings of the 3rd European Congress on Sign Language Research*, pages 355–379.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020a. Adversarial Training for Multi-Channel Sign Language Production. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2020b. Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021a. Continuous 3D Multi-Channel Sign Language Production via Progressive Transformers and Mixture Density Networks. In *International Journal of Computer Vision (IJCV)*.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2021b. Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1919–1929.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Thad Starner, Sean Forbes, Matthew So, David Martin, Rohit Sridhar, Gururaj Deshpande, Sam S. Sepah, Sahir Shahryar, Khushi Bhardwaj, Tyler Kwok, Daksh Sehgal, Saad Hassan, Bill Neubauer, Sofia Anandi Vempala, Alec Tan, Jocelyn Heath, Unnathi Kumar, Priyanka Mosur, Tavenner Hall, Rajandeep Singh, Christopher Cui, Glenn Cameron, Sohier Dane, and Garrett Tanzer. 2023. Popsign asl v1.0: An isolated american sign language dataset collected via smartphones. In *Neural Information Processing Systems*.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2018. Sign language production using neural machine translation and generative adversarial networks. In *BMVC*, volume 2019, pages 1–12.

Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. 2020. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Neha Tarigopula, Preyas Garg, Skanda Muralidhar, Sandrine Tornay, Dinesh Babu Jayagopi, and Mathew Magimai.-Doss. 2024. Content-based objective evaluation of artificially generated sign language videos. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3815–3819.

Neha Tarigopula, Sandrine Tornay, Ozge Mercanoglu Sincan, Richard Bowden, and Mathew Magimai Doss. 2025. Posterior-based analysis of spatio-temporal features for sign language assessment. *IEEE Open Journal of Signal Processing*.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. 2022. Motionclip: Exposing human motion generation to clip space. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 358–374. Springer.

Dave Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus. *Advances in Neural Information Processing Systems*, 36:29029–29047.

Jordan Voas, Yili Wang, Qixing Huang, and Raymond Mooney. 2023. What is the best automated metric for text to motion generation? In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11.

Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. 2024. T2S-GPT: Dynamic vector quantization for autoregressive sign language production from text. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3345–3356, Bangkok, Thailand. Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including signed languages in natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Zhengdi Yu, Shaoli Huang, Yongkang Cheng, and Tolga Birdal. 2024. Signavatars: A large-scale 3d sign language holistic motion dataset and benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–19.

Zhao Yuan, Zhang Ruiquan, Yao Dengfeng, and Chen Yidong. 2024. Translation quality evaluation of sign language avatar. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 405–415, Taiyuan,

China. Chinese Information Processing Society of China.

Biao Zhang, Garrett Tanzer, and Orhan Firat. 2024a. Scaling sign language translation. *arXiv preprint arXiv:2407.11855*.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023. DUB: Discrete unit back-translation for speech translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7147–7164, Toronto, Canada. Association for Computational Linguistics.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2024b. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. 2023. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20871–20881.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. Neural machine translation methods for translating text to sign language glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

Terry Yue Zhuo, Qiongkai Xu, Xuanli He, and Trevor Cohn. 2023. Rethinking round-trip translation for machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 319–337, Toronto, Canada. Association for Computational Linguistics.

Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. 2024a. Signs as tokens: An autoregressive multilingual sign language generator. *arXiv preprint arXiv:2411.17799*.

Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. 2024b. A simple baseline for spoken language to sign language translation with 3d avatars. In *European Conference on Computer Vision*, pages 36–54. Springer.

A Ham2Pose Metrics Re-Implementation via `pose-evaluation`

`pose-evaluation` toolkit enables the flexible creation of various metrics and pose processing pipelines. We successfully re-implemented the *nMSE*, *nAPE*, and *nDTW-MJE* metrics, and verified that the new implementation gave exactly identical results on a small collection of test files.

All metrics share certain preprocessing steps before comparison, in this order:

1. Remove `world landmarks`.
2. Calling the `reduce_holistic` function from the `pose-format` library, effectively reducing the keypoints to the face contour and the upper body.
3. Normalization by shoulder joints.
4. Hide low-confidence joint predictions.

In addition, *nMSE* and *nAPE* metrics do trajectory-based preprocessing, for each pair of keypoint trajectories:

1. Zero-pad the shorter trajectory.
2. Fill with zeros anywhere where either one of the trajectories is missing a value. For example, if trajectory A had [7, —, 7] and trajectory B had [—,8,8], the result would be thus: Trajectory A: [0,0,7], B: [0,0,8].

In contrast, the metrics implemented for the automatic meta-evaluation in §4, e.g., `DTW+Trim+MaskFill10.0+Hands-Only`, behave differently, filling each pose in without regard to the other. The result of trajectory A = [7,—,7] vs Trajectory B [—,8,8] would thus become A = [7,10,7] vs Trajectory B [10,8,8].

B Extended Human Evaluation Details

Figure 4 presents a screenshot of the Appraise platform we customized for the text-to-pose evaluation, where the instruction text is translated into English. The sign language versions of the instructions are linked: DSGS, LSF, LIS. The original text instructions in German, French, and Italian are below:

German Unten sehen Sie 10 Sätzen auf Deutsch (linke Spalten) und die entsprechenden möglichen Übersetzungen in Deutschschweizer Gebärdensprache (DSGS) (rechte Spalten). Bewerten Sie jede mögliche Übersetzung des Satzes. Sie können bereits bewertete Sätze jederzeit durch Anklicken eines Quelltextes erneut aufrufen und die Bewertung aktualisieren.

Bewerten Sie die Übersetzungsqualität auf einer kontinuierlichen Skala mit Hilfe der nachfolgend beschriebenen Qualitätsstufen:

0: Unsinn/Bedeutung nicht erhalten: Fast alle Informationen zwischen Übersetzung und Ausgangstext sind verloren gegangen. Es ist irrelevant, ob die Bewegungen natürlich sind.

2: Ein Teil der Bedeutung ist erhalten: Die Übersetzung behält einen Teil der Bedeutung der Quelle bei, lässt aber wichtige Teile aus. Die Erzählung ist aufgrund von grundlegenden Fehlern schwer zu verstehen. Bewegungen können mangelhaft sein.

4: Der grösste Teil der Bedeutung ist erhalten und die Bewegungen sind akzeptabel: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann kleine Fehler oder kleinere kontextuelle Unstimmigkeiten aufweisen. Bewegungen sehen teilweise nicht natürlich aus.

6: Perfekte Bedeutung und Natürlichkeit: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Bewegungen wirken natürlich.

French Vous voyez ci-dessous un document avec 10 phrases en français (colonnes de gauche) et leurs traductions candidates correspondantes langue des signes française (LSF) (colonnes de droite). Veuillez attribuer un score à chaque traduction possible de la phrase dans le contexte du document. Vous pouvez revisiter les phrases déjà évaluées et mettre à jour leurs scores à tout moment en cliquant sur un texte source.

Évaluez la qualité de la traduction sur une échelle continue en utilisant les niveaux de qualité décrits ci-dessous:

0: Absence de sens/aucune signification préservée: Presque toutes les informations sont perdues entre la traduction et la source. Le caractère naturel du mouvement n'est pas pertinent.

2: Une partie du sens est préservée: La traduction préserve une partie du sens de la source mais omet des parties importantes. Le récit est difficile à suivre en raison d'erreurs fondamentales. Le mouvement n'est pas toujours naturel.

4: La majeure partie du sens est préservée et le caractère naturel du mouvement est acceptable: La traduction conserve la majeure partie du sens de la source. Elle peut comporter quelques erreurs

mineures ou des incohérences contextuelles. Le mouvement peut sembler peu naturel.

6: Sens parfait et mouvements naturels: Le sens de la traduction est totalement cohérent avec la source et le contexte environnant (le cas échéant). Les mouvements sont naturels.

Italian Qui sotto trovate un documento con 10 frasi in italiano (colonne di sinistra) e lecorrispondenti possibili traduzioni nella lingua dei segni italiana (LIS) (colonne di destra). Valutate ogni possibile traduzione della frase nel contesto del documento. Potete rivedere le frasi valutate in precedenza e aggiornarne le valutazioni in qualsiasi momento cliccando sul testo sorgente.

Valutate la qualità della traduzione su una scala continua utilizzando i livelli di qualità descritti di seguito:

0: Privo di senso/significato non conservato: Quasi tutte le informazioni tra la traduzione e il testo sorgente sono andate perse. La naturalezza del movimento è inconsistente.

2: Parte del significato è conservato: La traduzione conserva parte del significato del testo sorgente, ma omette parti importanti. La narrazione è difficile da capire a causa di errori fondamentali. La naturalezza del movimento può essere insufficiente.

4: La maggior parte del significato è conservato e il movimento è accettabile: La traduzione conserva la maggior parte del significato del testo sorgente. Può contenere errori o discrepanze contestuali di entità minore. Il movimento può sembrare innaturale.

6: Significato perfetto e naturalezza: Il significato della traduzione è completamente coerente con il testo sorgente e con il contesto dato (se applicabile). Il movimento sembra naturale.

Figure 4: A screenshot of an example text-to-pose evaluation task in Appraise featuring sentence-level source-based direct assessment with custom annotator guidelines in German/French/Italian and DSGS/LSF/LIS, translated into English for readers' convenience.

Context is Ubiquitous, but Rarely Changes Judgments: Revisiting Document-Level MT Evaluation

Ahrii Kim
AI-Bio Convergence Research Institute
South Korea
ahriikim@gmail.com
trotacodigos/H-FALCON.git

Abstract

As sentence-level performance in modern Machine Translation (MT) has plateaued, reliable document-level evaluation is increasingly needed. While the recent FALCON framework with pragmatic features offers a promising direction, its reliability and reproducibility are unclear. We address this gap through human evaluation, analyzing sources of low inter-annotator agreement and identifying key factors. Based on these findings, we introduce **H-FALCON**, a **H**uman-centered refinement of FALCON. Our experiments show that, even with limited annotator consensus, H-FALCON achieves correlations comparable to or better than standard sentence-level protocols.

Furthermore, we find that contextual information is inherent in all sentences, challenging the view that only some require it. This suggests that prior estimates such as "n% of sentences require context" may stem from methodological artifacts. At the same time, we show that while context is pervasive, not all of it directly influences human judgment.

1 Introduction

The conventional approach to automatic machine translation (MT) evaluation has focused primarily on sentence-level analysis, emphasizing lexical overlap or n-gram similarity, as seen in BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015). More recent methods account for semantic similarity through embedding-based metrics such as BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020), while LLM-based (large language model) metrics, including XCOMET (Guerreiro et al., 2024) and Meta-Metrics (Anugraha et al., 2024), demonstrate improved alignment with human judgments. Despite these advances, their scope remains confined to sentence-level evaluation, failing to capture discourse phenomena such as cohesion, coreference, consistency, and pragmatic adequacy. Document-level metrics have been proposed (Jwalapuram et al. 2021; Zhao et al. 2023; Jiang et al. 2022), but they typically target narrow aspects of discourse and lack comprehensive coverage.

Human evaluation at the document level poses additional challenges due to the complexity of quantifying context-dependent phenomena. Approaches that rely only on overt discourse markers risk underestimating the role of context (Voita et al. 2019; Castilho 2022). Furthermore, protocols vary in context length, annotation granularity, and guideline specificity (Hardmeier et al. 2015; Kocmi et al. 2022). The resulting cognitive burden on evaluators can lead to longer annotation times and reduced inter-annotator agreement (IAA) (Läubli et al., 2018; Bawden et al., 2018; Graham et al., 2017). Collectively, these factors render document-level evaluation both methodologically complex and resource-intensive, limiting its adoption in MT research and practice (Sharma and Sridhar, 2025).

To address this gap, the FALCON framework (**F**unctional **A**ssessment of **L**anguage and **Co**ntextuality in **N**arratives; Kim 2025) integrates pragmatic features into a structured document-level protocol, with LLMs as judges. However, its human evaluation component remains underdeveloped and untested for reproducibility and reliability. We therefore conduct a meta-evaluation of FALCON through human assessments with professional translators, and extend the protocol by introducing H-FALCON, a reproducible and streamlined human evaluation framework. Our contributions are as follows:

- Conduct the first systematic reliability study of FALCON, identifying sources of inter-annotator variation,
- Provide a comprehensive meta-evaluation of FALCON across diverse proprietary models, revealing its limitations,

Proceedings of the Tenth Conference on Machine Translation, pages 81–97
November 8-9, 2025

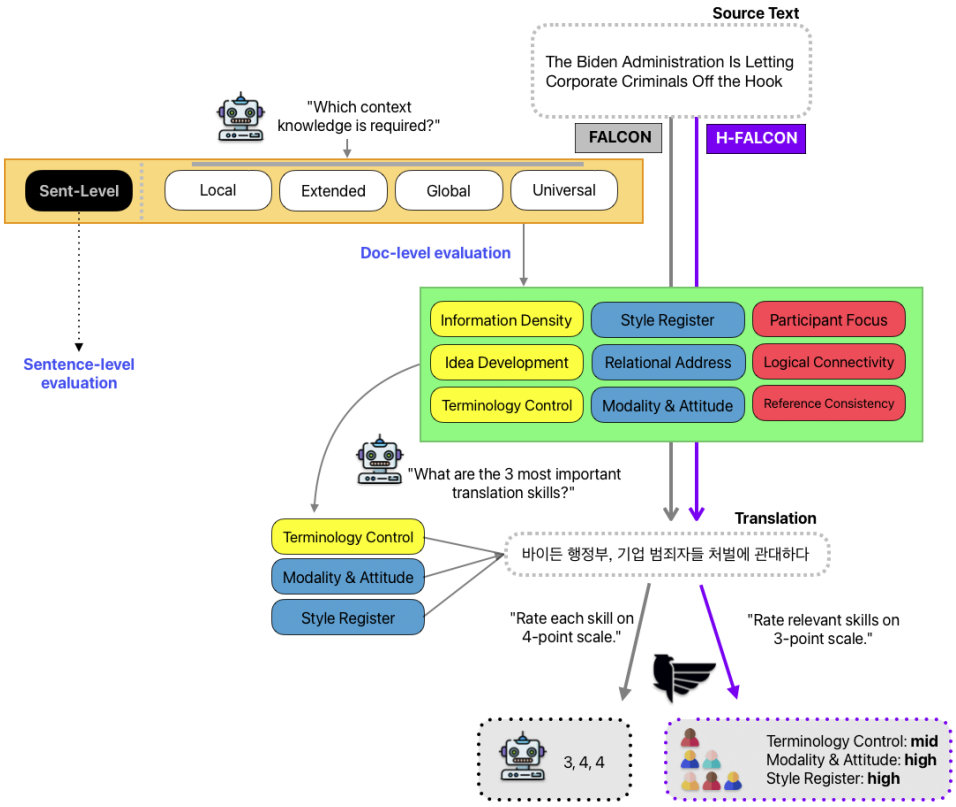


Figure 1: The evaluation process of FALCON consisting of labeling 1) relevant context knowledge and 2) assessment of translation skills, accompanied by 3) rating. This dual-phase process is integrated in H-FALCON by simultaneously conducting labeling and rating for all sentences.

- Introduce H-FALCON, a simplified and reliable protocol for document-level human evaluation,
- Present evidence that contextual information is inherent in all sentences,
- Demonstrate statistically that document-level evaluation contributes 10% to holistic evaluation scores.

2 Related Works

Document-level evaluation is not merely a scaled-up version of sentence-level evaluation; it captures translation phenomena that rely on extended context, such as coreference resolution, lexical cohesion, discourse connectives, and pragmatic intent (Thai et al. 2022; Dahan et al. 2024). These features recur across the document, with evidence distributed over multiple segments, shaping a distinctive atmosphere or nuance (Halliday and Matthiessen, 2004). Evaluating such phenomena enables a more accurate assessment of MT systems that appear statistically indistinguishable at the sentence level (Sharma and Sridhar, 2025). This section reviews prior efforts in both manual (§ 2.1) and automatic (§ 2.2) evaluation of document-level phenomena, including FALCON (§ 2.3).

2.1 For Manual Evaluation

The most visited sentence-level evaluation frameworks are MQM (Multidimensional Quality Metrics; Lommel et al. 2014) and TAUS DQF (Dynamic Quality Framework; Valli 2015). Their comprehensive error categories encompass some discourse elements such as *Language register* and *Inconsistent use of terminology*,[1], but predominantly focus on textual quality.

Document-level evaluation was initially driven by community efforts such as DiscoMT (Workshop on Discourse in Machine Translation; Hardmeier et al. 2015) and WMT (Conference on Machine Translation). Barrault et al. (2019) proposed a document-level scoring protocol (DR+DC), but its effectiveness was limited by low statistical power, often producing tied rankings. As a result, evaluations were shifted to the sentence level, either by considering adjacent segments (SR+DC) (Barrault et al., 2019) or entire documents (SR+FD)

[1] https://themqm.org/the-mqm-full-typology/

(Akhbardeh et al., 2021) to assess cross-sentence dependencies. The SR+DC approach later became standard practice (Kocmi et al. 2022; Kocmi et al. 2023), with Kocmi et al. (2024) extending the context window to ten consecutive sentences. In parallel, new error categories were introduced for discourse-related issues such as *Accuracy/Gender mismatch* and *Style/Archaic or obscure word choice* (Freitag et al., 2024). While these initiatives primarily focus on contextual conveyance, **our work broadens error typology by shifting from textual to discourse-level quality, systematically incorporating pragmatic, referential, and thematic dimensions into a structured protocol.**

2.2 For Automatic Evaluation

On the machine side, several automatic metrics have been developed to better capture discourse and context in MT evaluation. DiscoScore (Zhao et al., 2023) explicitly models discourse relations and coreference chains to assess cohesion and coherence. BlonDE (Jiang et al., 2022) integrates lexical, syntactic, semantic, and discourse-level features, making it suitable for narrative and dialogic text. Doc-COMET (Vernikos et al., 2022) extends COMET (Rei et al., 2020) to accept document-level inputs, leveraging contextual embeddings to evaluate translations within their broader discourse environment. While these approaches mark progress toward automated document-level evaluation, they generally emphasize only one or two discourse aspects—such as coherence or coreference—rather than offering a comprehensive, structured assessment of discourse phenomena.

Another line of research has focused on test suites targeting specific discourse elements. These include domain-specific investigations (Vojtěchová et al. 2019; Biçici 2019; Mukherjee and Yadav 2024; Bhattacharjee et al. 2024; Rozanov et al. 2024; Bawden and Sagot 2023), studies examining linguistic features (Avramidis et al. 2019; Popović 2019; Raganato et al. 2019; Zouhar et al. 2020; Macketanz et al. 2021; Manakhimova et al. 2023; Savoldi et al. 2023; Ármannsson et al. 2024; Friðriksdóttir 2024; Manakhimova and Macketanz 2024; Dawkins et al. 2024), and analyses incorporating discourse phenomena (Rysová et al. 2019; Kocmi et al. 2020; Avramidis et al. 2020; Scherrer et al. 2020; Mukherjee and Shrivastava 2023). DiscoBench (Wang et al., 2023) further addresses discourse-sensitive content, detecting pronoun mistranslation, topic drift, and other cross-sentence errors overlooked by sentence-level metrics.

Overall, these benchmarks highlight that document-level evaluation introduces qualitatively distinct challenges and opportunities, necessitating dedicated protocols and models for holistic MT quality assessment.

2.3 The FALCON Framework

FALCON (Functional Assessment of Language and Contextuality in Narratives; Kim 2025) proposes a structured protocol for document-level MT evaluation by incorporating pragmatic and discourse-level factors into a unified scoring scheme. It rests on two hypotheses:

(a) Document-level evaluation can be approximated at the sentence level if contextual information is effectively propagated across sentences.

(b) Such information can be inferred solely from the source, independent of the target language.

Discourse phenomena are classified into three meta-categories (MODE, TENOR, FIELD) and nine subcategories (specified in §3) collectively termed "translation skills." For each sentence, the judge selects the three most salient skills, with this restriction enhancing scoring stability.

Sentences not requiring context are first excluded through a labeling step, where annotators assign one of five context types (specified in §3). In the subsequent rating stage, each selected skill receives a 4-point score, as illustrated in Figure 1. Scores are then aggregated per segment or skill set to yield interpretable document-level indicators.

This protocol has so far been validated only indirectly: human annotators were asked to judge whether the model' s selections were appropriate, yielding an acceptance rate of 80.4% for context labeling and 71.6% for skill selection. However, no direct evaluation from a classification perspective has been conducted, which is the focus of the present study. Additional concerns may arise from the way context is presented, but this issue falls outside the scope of our work.

3 Experiment Setup

We conduct direct human evaluation of FALCON across two tasks and assess whether the current experimental design yields reproducible human judgments (§4). Using these gold annotations, we

Domain	Dataset	#Doc	#Seg	#Sent/Doc	#Sent/Seg
Canary	Original	1	1	–	–
	Ours	–	–	–	–
Literary	Original	8	206	74.13	2.88
	Ours	3	76	27.67	1.09
News	Original	17	149	19.53	2.23
	Ours	12	233	16.67	1.01
Social	Original	34	531	22.76	1.46
	Ours	23	500	23.26	1.07
Speech	Original	111	111	6.49	6.49
	Ours	–	–	–	–
All	Original	171	998	30.73	3.27
	Ours	38	809	22.53	1.06

Table 1: Comparison of dataset statistics between the original WMT24++ corpus and our filtered dataset. Here, **Seg** denotes a segment (paragraph in WMT24++), and sentence counts are reported per document and segment.

further perform a meta-evaluation to validate the framework's reliability (§5). The tasks are:

Task I: Context Knowledge Judges assign one of five levels of contextual knowledge required for translation: Sentence-level, Local, Extended, Global, Universal.

Task II: Translation Skills Judges select the three most relevant skills from nine predefined categories: Information Density, Idea Development, Terminology Control, Style Register, Relational Address, Modality & Attitude, Reference Consistency, Participant Focus, Logical Connectivity.

3.1 Dataset

The original WMT24++ English–Korean dataset (Deutsch et al., 2025) contains 998 segments from four domains (social, news, speech, and literary) with translations from ten systems. Because many segments span multiple sentences, it is unsuitable for our *sentence-level* design.

We construct a filtered subset while preserving domain balance. The speech domain is removed, as each document corresponds to a single segment without context, and the literary domain is partially pruned due to disproportionate length. Sentences with hyperlinks, hashtags, or timestamps are discarded, while emojis and user tags are retained as they are considered relevant to the evaluation.

The remaining data are re-segmented into individual sentences using NLTK (Bird et al., 2009) for English and KSS[2] for Korean. Source, target, and reference segments are automatically aligned with newline markers and then manually verified. For translation, we select the best-performing system (based on COMET scores), assuming that context-aware translation is unlikely from low-quality systems.

The final evaluation set consists of 809 unique sentences across three domains (social, news, and literary), preserving proportional domain distribution (Table 1). All retained segments preserve document boundaries, with sentence order tracked by custom IDs.

3.2 Recruitment & Training

We recruited three professional translators, all native Korean speakers with 5–10 years of English translation experience. For confidentiality, they were anonymized as Judge 1, Judge 2, and Judge 3 and are collectively referred to as judges. Based on the reduced segment length relative to the original dataset, we estimated an average throughput of 60 sentences per hour, corresponding to 13.3 hours per task and 27 hours in total per judge across two tasks. Judges were compensated at $30 per hour.

An online orientation was conducted via Google Meet to introduce the evaluation guidelines and demonstrate the platform. During the session, participants performed a preliminary evaluation using the platform. For the main study, judges were given one week to complete their evaluations. Time was tracked per item, and participants were instructed to maintain focus during annotation. They were provided full access to the document and permitted to review and revise their annotations prior to final submission.

3.3 Platform & Interface

We used Label Studio[3] as the evaluation platform (see Figure 9 in the Appendix). The interface allowed evaluators to consult label definitions, prior annotations, and relevant domain information throughout the task.

3.4 Metrics

We use IAA as the primary metric of reproducibility. For Task I, Cohen's Kappa (κ) is computed for each judge as in Equation 1, where P_o denotes the observed agreement and P_e the expected agreement under chance.

[2] https://github.com/hyunwoongko/kss
[3] https://labelstud.io

	(κ) ↑	(J) ↑
Judge 2–Judge 3	0.4995	0.6098
Judge 1–Judge 2	0.3883	0.4629
Judge 3–Judge 1	0.3646	0.4529
Avg.	**0.4175**	**0.5085**

Table 2: Pairwise IAA scores for Task I (Cohen's Kappa) and Task II (Jaccard similarity).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

For Task II, which involves multi-label annotation, we use the Jaccard similarity J (Equation 2), where A and B are the label sets from two annotators. For qualitative assessment, we collect participant feedback via Google Sheets and conduct subsequent linguistic analysis.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

4 Result

4.1 Reproducibility

Table 2 reports IAA for the two tasks. Across tasks, judges reach a **fair to moderate level of agreement** according to empirical standards (Landis and Koch 1977; Zhang and Zhou 2014; Rajpurkar et al. 2016), with average scores of $\kappa = 0.42$ and $J = 0.51$, and maximum scores of $\kappa = 0.50$ and $J = 0.61$. Agreement is highest between Judge 2 and Judge 3, suggesting that Judge 1 applied different criteria.

4.2 Analysis

We analyze disagreement by computing the proportion of pairwise label mismatches per task. For each judge pair, we identify the labels on which they disagreed and calculate their distribution. As shown in Figure 2, the largest divergence arises in the SENTENCE-LEVEL and LOCAL categories, accounting for 39.7% and 36.4% of disagreements, respectively, between Judge 1 and Judge 2.

To further examine this confusion, we merge related labels and recompute IAA. As shown in Table 3, the primary source of disagreement across judges lies in distinguishing SENTENCE-LEVEL from LOCAL. Merging these categories increases agreement from $\kappa = 0.4995$ to $\kappa = 0.58$.

To better understand this ambiguity, we analyze qualitative feedback on the difficulty of distinguishing context-independent (SENTENCE-LEVEL) from context-dependent labels. A recurring theme is the treatment of pronouns. For example, when the English pronoun "it" is translated into an equivalent pronoun in Korean and judged correct, the label is typically SENTENCE-LEVEL. By contrast, if the same translation is considered inadequate—requiring explicit mention of the referent noun—the label shifts to LOCAL. An illustrative case is shown in Table 4. As Judge 2 noted, "the interpretation of a pronoun's referent also influences verb choice, and thus I categorize the sentence as LOCAL."

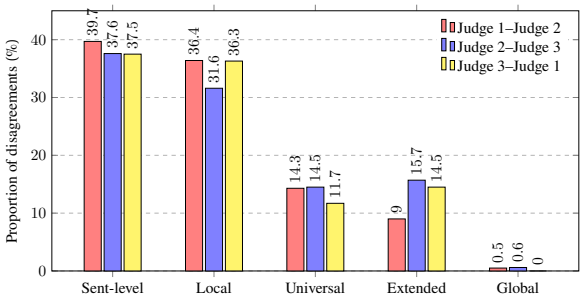


Figure 2: Distribution of Task I label disagreements across judge pairs (%).

Group	J_1–J_2 (Δ)	J_2–J_3 (Δ ↑)	J_3–J_1 (Δ)
L + S	**0.482 (+0.093)**	**0.580 (+0.080)**	**0.454 (+0.090)**
E + U	0.411 (+0.022)	0.568 (+0.069)	0.411 (+0.046)
E + L	0.414 (+0.026)	0.500 (+0.001)	0.397 (+0.033)
E + S	0.372 (–0.016)	0.480 (–0.020)	0.340 (–0.025)
L + U	0.372 (–0.017)	0.463 (–0.036)	0.343 (–0.022)
S + U	0.298 (–0.091)	0.418 (–0.081)	0.264 (–0.100)

Table 3: Cohen's κ after merging two labels. Parentheses indicate the change from the original κ. The highest agreement per column is shown in bold. Labels are abbreviated as L = Local, S = Sentence-level, E = Extended, and U = Universal. Judges are abbreviated as J_i.

SRC	I bought it like that and couldn't modify it, so I had to design around it.
TGT	구매했을 때부터 그런 형태였고, 수정할 수 없어서 그 형태에 맞춰 디자인해야 했어요.
BT	*It was in that form from the moment I purchased it, and since I couldn' t change it, I had to design everything to fit that shape.*

Table 4: A notable instance of pronoun provoking frequent misunderstanding between SENTENCE-LEVEL and LOCAL labels. The source (SRC) and target (TGT) segments are exemplified with the help of back-translation (BT).

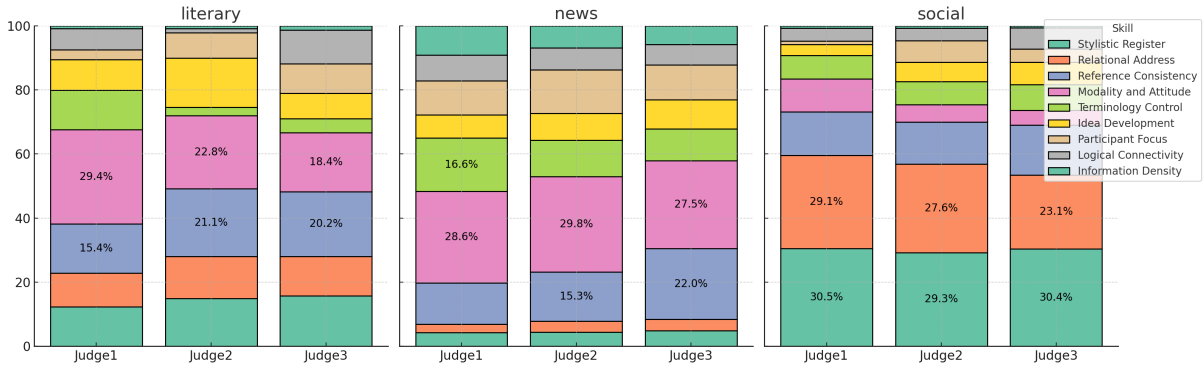


Figure 3: Distribution of Task II label choices across domains and judges. Values are shown for the largest slices.

4.3 Analysis

While sentence-level agreement on selected translation skills is limited, we analyze the distribution of skill choices per domain and per judge. Figure 3 shows that the three judges assign broadly similar proportions of skill labels across domains, suggesting that individual-level disagreements in exact label sets do not obscure shared evaluative priorities.

Closer inspection reveals domain-specific emphases. In the social domain, judges consistently highlight Style Register (avg. 30.1%) and Relational Address (26.6%), reflecting the importance of interpersonal stance in user-generated content. In the news domain, Modality & Attitude (28.6%) and Reference Consistency (16.7%) dominate, consistent with the demands for precision and coherence in reporting—a tendency also observed in literary text, where Modality & Attitude (23.5%) and Reference Consistency (18.9%) are most frequent. This indicates that low pairwise agreement does not necessarily reflect fundamental divergence, but rather differences in specific label selection. At the same time, the results point to a limitation of the current protocol: constraining annotators to exactly three skills per segment may not capture the full range of relevant judgments.

5 Meta-Evaluation of FALCON

The highest human IAA in our configuration is $\kappa = 0.50$ for Task I and $J = 0.61$ for Task II. Using these gold scores as reference, we evaluate the reliability of FALCON as an LLM-as-judge framework. As baselines, we test multiple proprietary models—OpenAI's `gpt-o3`, `o4-mini`, and the baseline from Kim (2025), `4.1-mini`. Model performance is assessed using the same reproducibility metrics defined in §3.4, complemented by accuracy for Task I, where the output is a single categorical label, and Micro-F1 for Task II, where multiple labels must be selected simultaneously.

Group	Pair	acc (%)↑	κ
vs.	**J_2, J_3**	**70.09**	**0.4995**
	J_1, J_2	66.25	0.3883
	J_1, J_3	62.92	0.3646
vs.	**J_3, `o4-mini`**	**53.89**	**0.2535**
	J_1, `o4-mini`	52.29	0.1788
	J_2, `o4-mini`	51.67	0.1891
	J_3, `o3`	51.17	0.2059
	J_1, `o3`	50.31	0.1484
	J_2, `o3`	49.57	0.1591
	J_1, `4.1-mini`	42.77	0.0802
	J_2, `4.1-mini`	39.80	0.0478
	J_3, `4.1-mini`	39.68	0.0750
vs.	**`o3`, `o4-mini`**	**71.69**	**0.5239**
	`4.1-mini`, `o4-mini`	47.22	0.2046
	`o3`, `4.1-mini`	40.30	0.1068

Table 5: Pairwise accuracy and Cohen's Kappa κ by human () and model () groups for Task I.

5.1 Reliability of context knowledge

Table 5 reports pairwise accuracy and IAA across human–human, human–model, and model–model comparisons. The best-performing model, `o4-mini`, achieves 53.89% accuracy, which falls short of even the weakest human pair (Judge 1–-Judge 3, 62.92%). No model approaches the agreement level of the strongest human pair (Judge 2–-Judge 3). The concordance with human annotations remains at most **fair** ($\kappa = 0.25$ for `o4-mini`), underscoring the limited ability of current LLMs to reliably distinguish context categories at a human-comparable level.

To better understand this gap, we analyze which labels drive model–human discrepancies. Figure 4

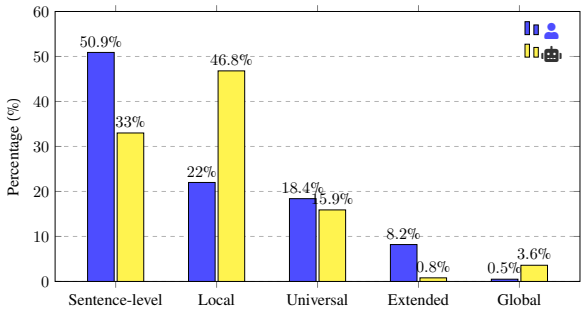


Figure 4: Task I label distribution of disagreements between Judge 2 ( ) and `o4-mini` ( ), identified as the most aligned human–model pair.

illustrates the disagreement distribution between Judge 2 and `o4-mini`, the pair with the highest human–model consensus. The largest share of divergence arises from SENTENCE-LEVEL (50.9%) from the human part and from LOCAL (46.8%) from the machine part. These categories mirror the main sources of confusion among human annotators, suggesting that while models replicate human-like weaknesses, they lack the robustness to resolve such ambiguities consistently.

5.2 Reliability of translation skill

Table 6 shows that the strongest human–model agreement is attained with `o4-mini` ($J = 0.406$), substantially lower than both human–human and model–model levels. Model precision reaches 53.6%, comparable to the earlier task, but still insufficient to approximate human reliability. Interestingly, model–model agreement is relatively high, reaching up to $J = 0.597$, on par with the stronger human–human pairs.

These findings suggest that models produce consistent predictions across systems, yet this consistency reflects shared internal heuristics rather than alignment with human reasoning. While human annotators converge through pragmatic interpretation, models seem to exploit surface-level patterns that do not fully capture evaluative criteria. Closing this gap demands not just higher accuracy, but agreement with humans based on human-like reasoning.

5.3 Summary

The central hypothesis of FALCON—that document-level evaluation can be approximated at the sentence level—requires caution. Our results show that judges often confuse adjacent levels of context, underscoring the need for clearer definitions

Group	Pair	avg. $J\uparrow$	f1
vs.	J_2–J_3	**0.6098**	**0.7183**
	J_1–J_2	0.4629	0.5915
	J_1–J_3	0.4529	0.5737
vs.	J_2, `o4-mini`	**0.4067**	**0.5360**
	J_3, `o4-mini`	0.3976	0.5272
	J_2, `o3`	0.3970	0.5231
	J_1, `o4-mini`	0.3912	0.5196
	J_1, `o3`	0.3829	0.5099
	J_2, `4.1-mini`	0.3704	0.4931
	J_3, `4.1-mini`	0.3683	0.4893
	J_1, `4.1-mini`	0.3660	0.4871
	J_3, `o3`	0.3625	0.4854
vs.	`o3`, `o4-mini`	**0.5972**	**0.7082**
	`4.1-mini`, `o4-mini`	0.4665	0.5948
	`4.1-mini`, `o3`	0.4250	0.5554

Table 6: Average pairwise Jaccard Similarity J and Micro F1 between human ( ) and model ( ) groups for Task II.

of "context." Furthermore, the low agreement in Task II suggests that identifying universal translation skills solely from the source text risks poor reproducibility of gold judgments.

6 Refined Protocol: H-FALCON

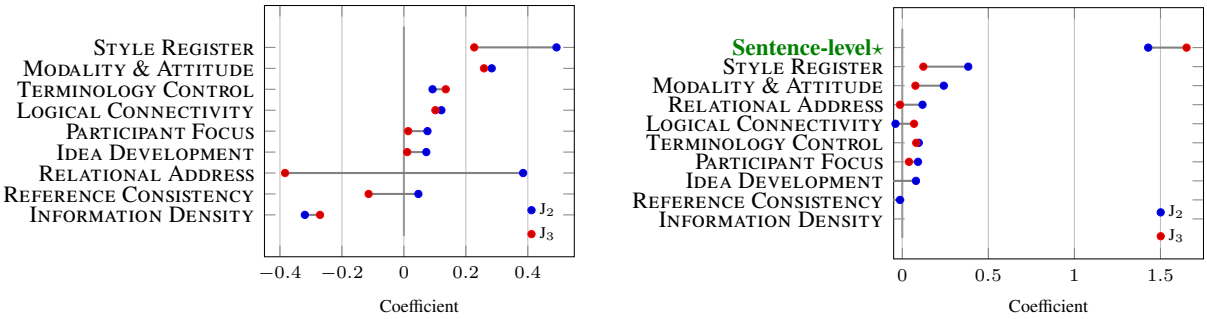
The current protocol of FALCON suffers from ambiguous definitions of context and limited reproducibility in skill selection, calling into question its central hypotheses. Building on these findings, we identify three structural limitations of FALCON: unclear translation objectives for human evaluators, the rigid requirement to assign exactly three skills per sentence, and the lack of adaptability to the domain and language pair.

To address them, we propose H-FALCON (**H**uman-centered FALCON), grounded in two revised hypotheses: (i) every sentence is influenced by context, and (ii) judges should flexibly decide the number of translation skills.

6.1 Design

Given these assumptions, H-FALCON removes Task I, since all sentences are subject to evaluation. For Task II, rather than selecting a fixed set of relevant skills, judges directly evaluate the pertinence of each skill, thereby unifying annotation and rating into a single step (Figure 1).

To support this protocol, every skill is initialized as NOT RELEVANT. Judges then assign one of three



(a) Doc Features

(b) Doc + Sent Features

Figure 5: Linear regression coefficients for Judge 2 (J_2) and Judge 3 (J_3) with (b) and without (a) **sentence-level** score. Features with scores near 0 have minimal influence on the holistic score.

ratings—High, Medium, or Low—following House's theoretical framework (House, 2015). This triadic scale replaces the 4-point scheme of Kim (2025), to represent preliminary feedback from our evaluators that three levels suffice, as discourse phenomena often lend themselves to relatively clear judgments.

6.2 Experiment

To verify the reproducibility of the refined H-FALCON protocol, we sample 300 new instances from WMT24++ (Deutsch et al., 2025) that are not included in the earlier experiments. Human evaluation is conducted by Judge 2 and Judge 3, the pair with the highest agreement in prior tasks.

The evaluation environment remains unchanged, using the same platform as in Figure 10. In this setting, judges simultaneously select and rate relevant skills, eliminating the separation of annotation and scoring. To provide additional baselines, we also collect MQM-style sentence-level error annotations on a 4-point scale and holistic quality scores (sentence + document level) on a 10-point scale. These parallel evaluations allow us to establish a benchmark IAA threshold for H-FALCON and to examine relationships among the three metrics. All ratings are obtained at the sentence level, and scale variation is deliberately employed to minimize task confusion.

The reliability of skill selection is measured by excluding Not Relevant labels and computing Jaccard similarity between the two judges. Correlations between evaluation metrics are quantified using Pearson, Spearman, and Kendall's tau coefficients.

6.3 Reproducibility of H-FALCON

The Jaccard similarity for overlapping translation skills between the two judges is 0.532, remaining consistently low and consistent with the earlier experiment. This highlights the inherent difficulty of achieving consensus, regardless of the method of label collection.

To further examine how the judges weigh each skill when assigning holistic scores, we fit a linear regression model for each judge, using the holistic score as the dependent variable and the individual label scores as predictors (with an intercept). This analysis quantifies the relative contribution of each skill while controlling for the others. As shown in Figure 5 (a), the judges diverge most clearly on Relational Address: Judge 2 associates higher holistic scores with stronger performance in this skill, whereas Judge 3 tends to assign lower scores. A similar but weaker divergence is observed for Reference Consistency. Importantly, these opposite directions remain significant within 95% confidence intervals, underscoring that the divergence reflects genuine differences in evaluative criteria rather than statistical noise. These divergent patterns suggest that the guidelines for the labels may require refinement and additional evaluator training to ensure consistent application.

6.4 Further Analysis

H-FALCON score as a proxy measure

We examine whether the obtained labels can serve as proxies for document-level scoring. Each annotation is assigned a numerical value (High=3, Medium=2, Low=1, Not Relevant=0), and scores are computed either by aggregating values ("sum") or by counting non-zero labels. Correlation between the two judges across sentence-, document-, and holistic-level scores (Table 7) indicates that the document-level scheme achieves agreement comparable to sentence-level evaluation ($\rho = 0.55$ vs.

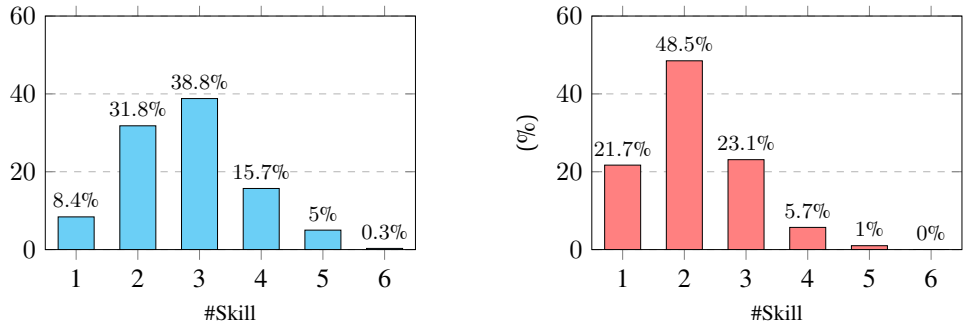


Figure 6: Distribution of the number of selected skills per sentence for each judge (left: Judge 2, right: Judge 3).

0.44). Notably, the counting method yields slightly higher consensus (0.55 vs. 0.48), highlighting its potential as an effective approach for annotating document-level quality.

Type	Pearson	Spearman	Kendall
Sentence-level	0.494	0.441	0.413
H-FALCON (sum)	0.499	0.483	0.378
H-FALCON (count)	0.562	0.545	0.486
Holistic	0.650	0.587	0.502

Table 7: Correlations between two raters across sentence-level, document-level (using two aggregation styles), and holistic scores.

Limited explanatory power of document-level score

To further assess the relative impact of sentence- and document-level features on holistic judgments, we extend the regression model by adding the sentence-level score as an independent variable. As shown in Figure 5 (b), the sentence-level score is the strongest predictor of holistic quality, with coefficients of 1.43 (95% CI: 1.22–1.63) for Judge 2 and 1.65 (95% CI: 1.49–1.82) for Judge 3.

Table 8 reports the explanatory power (R^2) of models with and without the sentence-level score. Document-level scores alone account for little variance in holistic judgments ($R^2 = 0.11$), explaining only 11% of the variance in holistic judgments. However, incorporating the sentence-level score increases explanatory power to 0.54 and reduces the intercept from 7.11 to 2.29. These results confirm that sentence-level quality is the primary driver of holistic assessments.

At least one discourse feature per sentence

We calculate the number of translation skills annotated per judge. Figure 6 shows that every sentence is annotated with at least one skill, most frequently with three to four skills (38.8% and 48.5% for Judge 2 and Judge 3, respectively). This finding challenges the claim that only a subset of sentences requires contextual information (Castilho, 2022). On the contrary, we emphasize that contextual information can influence translation in all cases—even for simple utterances such as "hi." However, as shown in the previous section, its impact on the holistic score is relatively limited. Still, this does not diminish the importance of document-level evaluation, which remains a key factor for distinguishing higher-performing models.

	Doc			Doc + Sent		
	J_2	J_3	Avg	J_2	J_3	Avg
R^2	0.12	0.09	0.11	0.47↑	0.61↑	0.54↑
Intercept	6.46	7.76	7.11	2.10↓	2.48↓	2.29↓

Table 8: The explanatory power (R^2) of models with document-level score (**Doc**) and with document- and sentence-level scores (**Doc+Sent**). Doc+Sent results are highlighted.

7 Conclusion

Our findings challenge prevailing assumptions in MT evaluation by demonstrating that contextual information, though modest in magnitude, is both universal and consequential for human judgment. Operationalizing this insight, H-FALCON provides a reproducible, context-aware evaluation protocol that aligns as closely with human preferences as traditional sentence-level approaches. These results underscore the need to move beyond narrow, sentence-bounded metrics toward richer document-level assessments that capture the pragmatic realities of translation quality. As MT performance converges at the sentence level, such holistic, context-sensitive evaluation will be essential for driving the next phase of progress in the field.

8 Limitation

Our study is limited to a single mid-resourced language pair. While this is acceptable given our focus on the human evaluation setting—which is largely consistent across languages—the reproducibility and reliability of FALCON may be underestimated. For the same reason, we did not experiment with other open-weight models such as LLaMA or Mistral.

On the human side, only three annotators were engaged, one of whom showed notably divergent behavior. In addition, even under the refined protocol, the consensus on translation skills remained low (§ 6.3). These issues highlight the need for more proactive calibration sessions among annotators.

Finally, we did not investigate how context should be presented or which types of context were most informative on the target side for FALCON. We leave this as an avenue for future work.

9 Acknowledgment

This research was supported by G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317).

References

Farhad Akhbardeh, Andrey Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Christian Federmann, Yvette Graham, Barry Haddow, Kenneth Heafield, Philipp Koehn, Christof Monz, and Others. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation* (*WMT21*), pages 1–88. Association for Computational Linguistics.

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. 2024. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Aljoscha Burchardt, and et al. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. *arXiv preprint arXiv:2010.06359*.

Eleftherios Avramidis, Vivien Macketanz, Ulrich Strohriegel, and Aljoscha Burchardt. 2019. Linguistic evaluation of german-english machine translation using a test suite. *arXiv preprint arXiv:1910.07457*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Barry Haddow, Chris Hokamp, Philipp Koehn, Shervin Malmasi, Christof Monz, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation* (*Volume 2*: *Shared Task Papers*, *Day 1*), pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume 1* (*Long Papers*), pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Soham Bhattacharjee, Biswajit Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in english-hindi translation. In *Proceedings of the Ninth Conference on Machine Translation* (*WMT24*). Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python*: *Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Ergun Biçici. 2019. Machine translation with parfda, moses, kenlm, nplm, and pro. In *Proceedings of the Fourth Conference on Machine Translation* (*Volume 2*: *Shared Task Papers*, *Day 1*), pages 66–73. Association for Computational Linguistics.

Sheila Castilho. 2022. How much context span is enough? examining context-related issues for document-level MT. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3017–3025, Marseille, France. European Language Resources Association.

Nicolas Dahan, Rachel Bawden, and François Yvon. 2024. Survey of automatic metrics for evaluating machine translation at the document level. Technical report, HAL Open Science. Available at HAL Open Science.

Hillary Dawkins, Isar Nejadgholi, and Chi-Kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleni Briakou, Isaac Caswell, Max Finkelstein, Roni Galor, and 1 others. 2025. WMT24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Sigríður Rut Friðriksdóttir. 2024. The genderqueer test suite. In *Proceedings of the Ninth Conference on Machine Translation (WMT24)*, pages 265–273. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Michael Alexander Kirkwood Halliday and Christian Matthias Ingemar Martin Matthiessen. 2004. *An Introduction to Functional Grammar*, 3rd edition. Hodder Arnold.

Christian Hardmeier, Liane Guillou, Pierre Lison, and Jörg Tiedemann. 2015. Report on the discomt 2015 shared task on discourse translation. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 1–16. ACL.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge, London and New York.

Zheng Jiang, Yang Yu, Yang Feng, Bing Qin, and Ting Liu. 2022. Blonde: An automatic evaluation metric for document-level natural language generation. In *Proceedings of NAACL*, pages 1679–1698.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. 2021. Dip benchmark tests: Evaluation benchmarks for discourse phenomena in {mt}.

Ahrii Kim. 2025. Falcon: Holistic framework for document-level machine translation evaluation. *TechRxiv*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, Maja Popović, and Mariya Shmatova. 2022. Findings of the 2022 conference on machine translation (wmt22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45, Abu Dhabi. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. *arXiv preprint arXiv:2010.06018*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for defining translation quality. In *Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'14)*, pages 1285–1291. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, and Aljoscha Burchardt. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for german to english and english to german. In *Proceedings of the Sixth Conference on Machine Translation* (*WMT21*), pages 1122–1137. Association for Computational Linguistics.

Sabina Manakhimova, Eleftherios Avramidis, and Vivien Macketanz. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can chatgpt outperform nmt? In *Proceedings of the Eighth Conference on Machine Translation* (*WMT23*). Association for Computational Linguistics.

Sabina Manakhimova and Vivien Macketanz. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation* (*WMT24*). Association for Computational Linguistics.

Anwesha Mukherjee and Manish Shrivastava. 2023. Iiit hyd's submission for wmt23 test-suite task. In *Proceedings of the Eighth Conference on Machine Translation* (*WMT23*). Association for Computational Linguistics.

Anwesha Mukherjee and Shruti Yadav. 2024. Cost of breaking the llms. In *Proceedings of the Ninth Conference on Machine Translation* (*WMT24*). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2019. Evaluating conjunction disambiguation on english-to-german and french-to-german wmt 2019 translation hypotheses. In *Proceedings of the Fourth Conference on Machine Translation* (*Volume 3*: *Shared Task Papers*, *Day 2*), pages 597–602. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The mucow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation* (*Volume 3*: *Shared Task Papers*, *Day 2*), pages 603–611. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 2383–2392. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pages 2685–2702, Online. Association for Computational Linguistics.

Nikita Rozanov, Vladislav Pankov, and Danila Mukhutdinov. 2024. Isochronometer: A simple and effective isochronic translation evaluation metric. *arXiv preprint arXiv:2410.11127*.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A test suite and manual evaluation of document-level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation* (*Volume 2*: *Shared Task Papers*, *Day 1*), pages 455–463, Florence, Italy. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in mt with must-she and ines. *arXiv preprint arXiv:2310.19345*.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The mucow word sense disambiguation test suite at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation* (*WMT20*).

Himanshu Sharma and Bharat Ram Sridhar. 2025. Document-level machine translation through discourse modelling: A survey. *CFILT - IITB*.

Katherine Thai, Magdalena Karpinska, Kalpesh Krishna, Baishakhi Ray, Kathleen McKeown, Ron Artstein, and Benjamin Van Durme. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1256–1274, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paola Valli. 2015. The TAUS quality dashboard. In *Proceedings of the 37th Conference Translating and the Computer*, pages 127—136, London, UK. AsLing.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation* (*WMT*), pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Tereza Vojtěchová, Matúš Novák, Matěj Klouček, and Ondřej Bojar. 2019. Sao wmt19 test suite: Machine translation of audit reports. *arXiv preprint arXiv:1909.01701*.

Longyue Wang, Zefeng Du, Donghuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, and Zhaopeng Tu. 2023. Disco-bench: A discourse-aware evaluation benchmark for language modelling. *Preprint*, arXiv:2307.08074.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26 (8):1819–1837.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wei Zhao, Michael Strube, and Steffen Eger. 2023. DiscoScore: Evaluating text generation with BERT and discourse coherence. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia. Association for Computational Linguistics.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. Wmt20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation* (*WMT20*), pages 347–356. Association for Computational Linguistics.

Björn Ármannsson, Hrafn Hafsteinsson, and Atli Jasonarson. 2024. Killing two flies with one stone: An attempt to break llms using english→icelandic idioms and proper names. *arXiv preprint arXiv:2410.03394*.

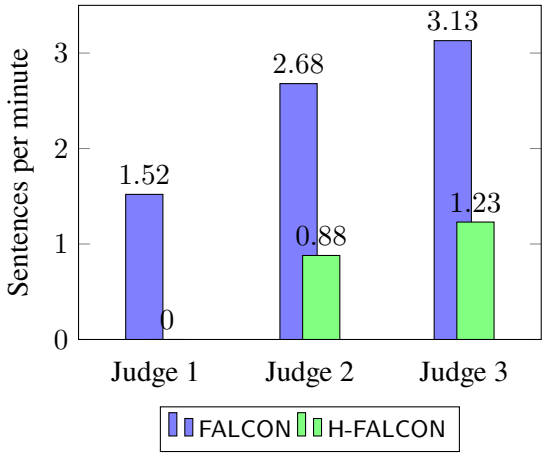


Figure 7: Average throughput per judge in FALCON vs. H-FALCON. Judge 1 was not hired for H-FALCON.

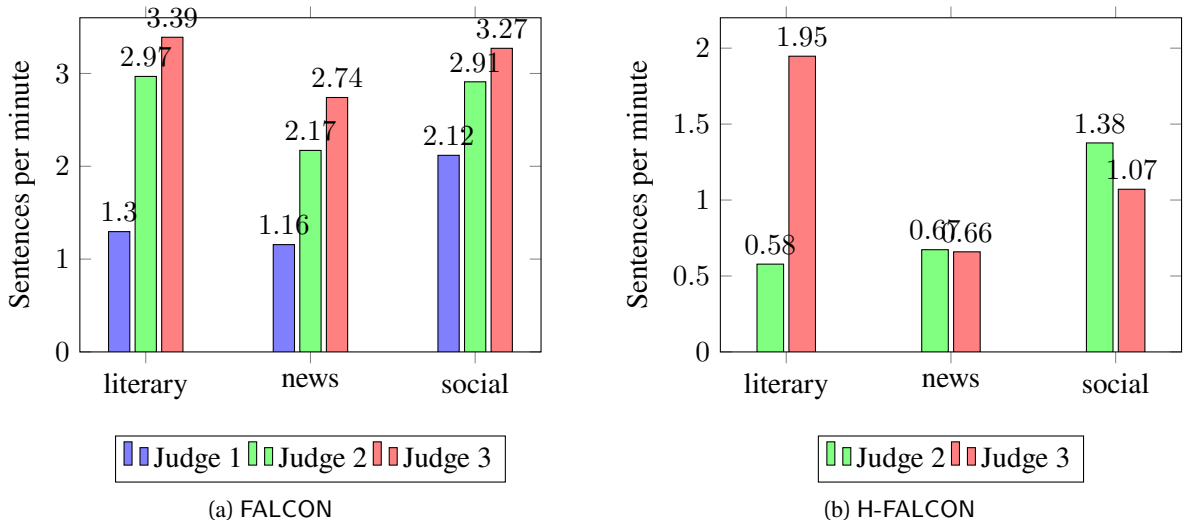
Appendix

A Evaluation Throughput

We calculate throughput per judge under two different frameworks: FALCON and H-FALCON. The key distinction is that the H-FALCON setting requires both label annotation and rating, which introduces additional cognitive load and time, whereas the FALCON condition measures throughput without the rating phase.

Figure 7 shows that throughput values are consistently lower in H-FALCON than in FALCON, reflecting the extra annotation steps. For example, the average throughput per judge decreases from 1.52–3.13 sent/min in FALCON to 0.88–1.23 sent/min in H-FALCON. This suggests that rating is the most time-consuming component of the evaluation: despite H-FALCON consolidating the task into a single step, throughput falls to less than half of FALCON, indicating that the rating phase dominates the overall processing time.

When examining domain-level performance in Figure 8, consistent patterns emerge across both setups. Social texts yield the highest throughput, reflecting their relatively simple and conversational style, while literary texts slow down judges the most, likely due to complex syntax and stylistic density. News texts fall in between, with moderate difficulty and processing speed. This ordering is preserved in both FALCON and H-FALCON, though absolute throughput values are lower in the latter due to the added annotation and rating tasks. These results confirm that genre characteristics strongly shape translation throughput, and that such effects remain robust even under heavier annotation requirements.



(a) FALCON (b) H-FALCON

Figure 8: Throughput by domain and judge across FALCON and H-FALCON setups. Higher values indicate faster processing.

B Descriptions of Context Levels

Sentence-level The sentence can be fully understood and translated without any outside information. All necessary meaning is present within the sentence itself — vocabulary, grammar, and semantics are straightforward.

Local Understanding requires minimal surrounding context — maybe the previous or next sentence — but nothing broader. Without it, pronouns, references, or logical connectors might be confusing.

Extended Grasping the meaning requires understanding the broader scene, paragraph, or emotional flow. Cultural nuance, emotional undertones, or evolving character perspectives start to matter.

Global The sentence depends on knowledge of the entire work (novel, article, movie) or even multiple entries (book series, TV seasons). Important world-building, character arcs, fictional history, or long-term motifs influence meaning.

Universal Understanding draws on extensive external knowledge — history, philosophy, science, mythology, social structures, or famous world events. Without that shared knowledge, translation risks misfiring badly.

C Descriptions of Translation Skills

Information Density Does the sentence compress information into abstract or complex structures required by the genre or audience? Important linguistic devices are nominalization, complex noun phrases, embedded clauses, compounding, metaphors, analogies, symbolic imagery, etc.

Idea Development Do some elements in the sentence influence the development of the central theme and the rhetorical structure expected by the genre? Important linguistic devices are discourse markers, schematic structures (e.g., introduction-body-conclusion), paragraph transitions, etc.

Terminology Control Does the sentence have technical or domain-specific vocabulary that requires accurate and consistent use across an entire text? Important linguistic devices are technical nouns, specialized terminology, standard collocations, fixed expressions, etc.

Style Register Do some elements in the sentence require a degree of linguistic politeness and stylistic appropriateness suited to the context and purpose of the text? Important linguistic devices are lexical choice, pronoun usage, verb conjugation, discourse markers, euphemisms, idiomatic expressions, etc.

Reference Consistency Does the sentence contain elements that refer to the same entity within

(a) Task I

Label	J_1	J_2	J_3	Avg↑
Sent-level	63.16	60.57	54.14	59.29
Local	23.11	21.76	26.33	23.73
Universal	10.88	13.23	10.01	11.37
Extended	2.84	4.08	9.52	5.48
Global	0.00	0.37	0.00	0.12

(b) Task II

Label	J_1	J_2	J_3	Avg↑
Style Register	21.26	20.77	21.67	21.23
Relational Address	19.70	19.28	16.44	18.47
Reference Consistency	13.51	14.50	17.84	15.28
Modality and Attitude	17.39	14.05	12.48	14.64
Terminology Control	10.47	7.95	8.24	8.89
Idea Development	5.07	7.62	7.70	6.80
Participant Focus	4.04	8.82	6.55	6.47
Logical Connectivity	5.40	4.49	6.84	5.58
Information Density	3.17	2.51	2.22	2.63

Table 9: Proportion of Task I, II labels annotated by three judges (%).

the text? The consistent use of such elements creates connections and coherence and ensures clear identification of participants, objects, and ideas throughout the text. Important linguistic devices are reference, substitution of clause, gender/tense/number agreement, deixis, ellipsis, repetition, synonyms, etc.

Logical Connectivity Does the sentence have connectors or structures that require clear expression of relationships — such as cause, contrast, or sequence — between ideas? Important linguistic devices are logical connectors (e.g., however, therefore), adversatives, causal linkers, etc.

Modality and Attitude Do some elements in the sentence express possibility, obligation, certainty, or speaker/writer's stance that convey the text's mood and tone? Important linguistic devices are modal verbs and auxiliaries (e.g., must, might), evaluative adjectives (e.g., important, unfortunate), stance adverbs (e.g., perhaps, clearly, surprisingly), emotionally charged expressions, subjunctive or conditional constructions, etc.

Relational Address Does the sentence rely on an understanding of the author's cultural, historical, or social background that affects his/her voice, intent, and the nuanced relationships with listener/reader? Important linguistic devices are gendered forms, titles and vocatives, pronoun, honorifics, relational expressions, sociolect, etc.

Participant Focus Should the emphasis of the sentence on key participants or elements (such as people, places, or objects) be preserved to convey the original meaning across a text? Important linguistic devices are subject-specific terminology, transitivity structures (verb types, selection of active/passive, selection of grammatical subject, use of nominalization instead of verb), etc.

D Analysis of Collected Data

Table 9-(a) reports the number of annotations per context type, indicating broadly consistent distributions across judges. Roughly 60% of sentences were judged as translatable without additional context, though the exact subset of sentences varied considerably by annotator. Among context-dependent categories, Local was the most frequent, averaging 24%. By contrast, Global was almost never selected, suggesting that this type of context is difficult to capture reliably at the sentence level.

Turning to translation skills in Table 9-(b), Style Register (21.23%) and Relational Address (18.47%) emerged as the most frequently required skills, aligning with qualitative feedback that highlights their importance in context-sensitive translation. Conversely, Information Density was rarely chosen (2.6%), which may reflect either limited judge awareness or the relatively low salience of this feature in the dataset. These observations underscore the need for further clarification of certain skill definitions to improve annotation reliability.

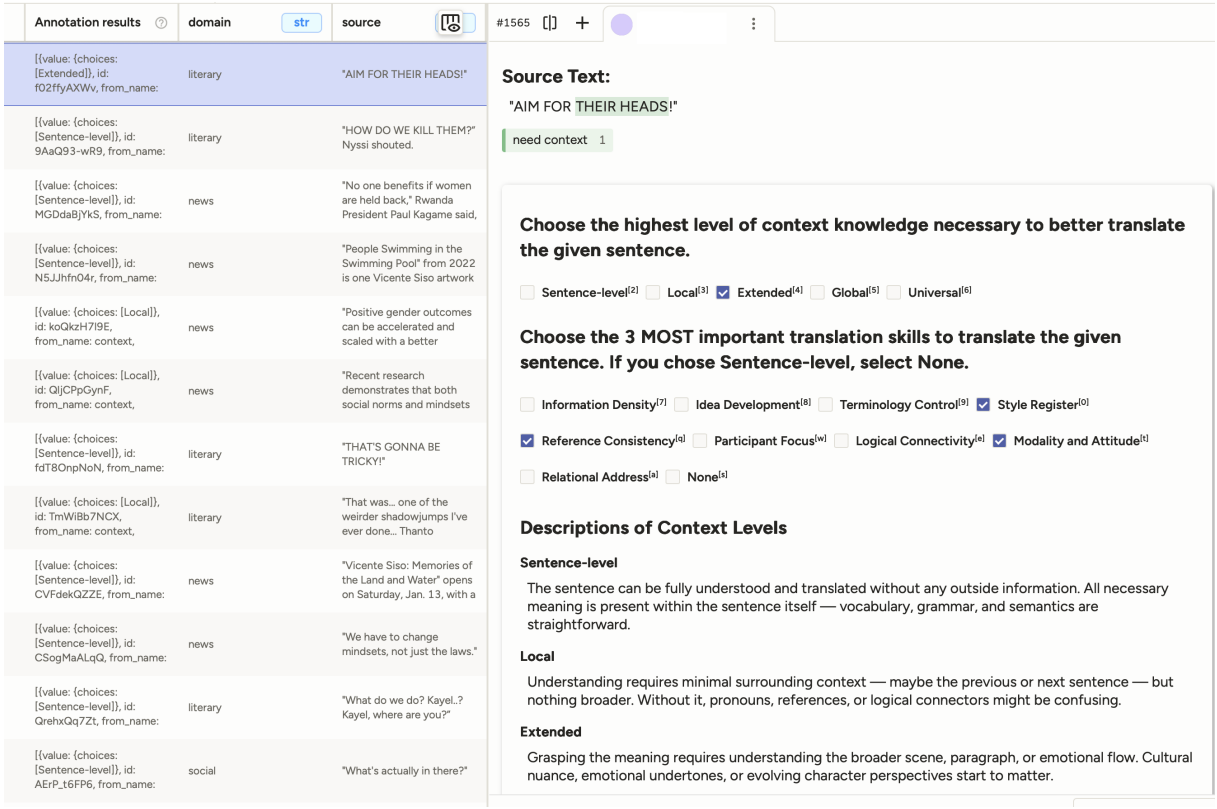


Figure 9: Label Studio interface for human evaluation in FALCON, showing labels of Task I and II. Expanded views provide consistent explanations for each category.

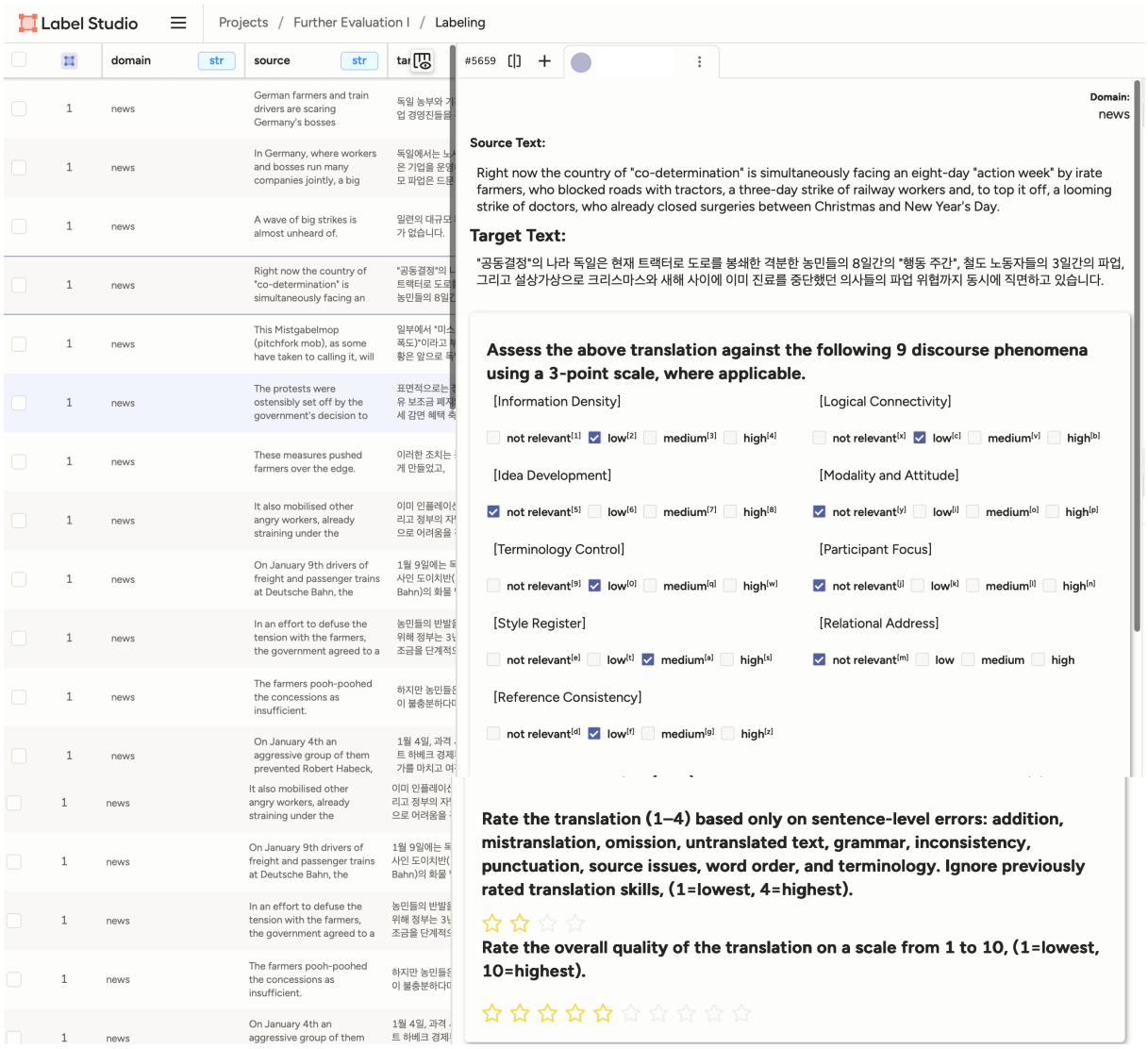


Figure 10: Interface of Label Studio for the human evaluation in H-FALCON. All translation skills are set to “not relevant” by default, and both sentence-level and holistic scores are collected concurrently.

GIIFT: Graph-guided Inductive Image-free Multimodal Machine Translation

Jiafeng Xiong
Department of Computer Science
University of Manchester
jiafeng.xiong@manchester.ac.uk

Yuting Zhao
Department of Advanced Information Technology
Kyushu University
zhao.yuting.095@m.kyushu-u.ac.jp

Abstract

Multimodal Machine Translation (MMT) has demonstrated the significant help of visual information in machine translation. However, existing MMT methods face challenges in leveraging the modality gap by enforcing rigid visual-linguistic alignment whilst being confined to inference within their trained multimodal domains. In this work, we construct novel multimodal scene graphs to preserve and integrate modality-specific information and introduce **GIIFT**, a two-stage **G**raph-guided **I**nductive **I**mage-**F**ree MM**T** framework that uses a cross-modal Graph Attention Network adapter to learn multimodal knowledge in a unified fused space and inductively generalize it to broader image-free translation domains. Experimental results on the Multi30K dataset of English-to-French, English-to-German, and English-to-Ukraine tasks demonstrate that our GIIFT surpasses existing MMT baselines even without images during inference. Results on the WMT benchmark show significant improvements over the image-free MMT translation baselines, demonstrating the strength of GIIFT towards inductive image-free inference [1].

1 Introduction

Multimodal machine translation (MMT) aims to improve traditional text-only neural machine translation (NMT) by incorporating multimodal data, particularly visual inputs (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Existing methods mostly focus on forcing the alignment between image and text to improve MMT, that they have demonstrated the effectiveness benefited from the aligned visual information in disambiguating text (Ive et al., 2019; Zhao et al., 2022b; Futeral et al., 2023). However, an aligned form of multimodal dataset in both the training and inference phases has disabled the MMT model to generalize further in the normal pure-text machine translation setup. Although the rich information of images can bring benefits to translation beyond the level of text, when aligned image information is indispensable for inference in the translation process, the application of MMT models will be severely limited. Therefore, the inability to get rid of aligned images in the inference phase is the critical bottleneck for the flexible application of MMT models.

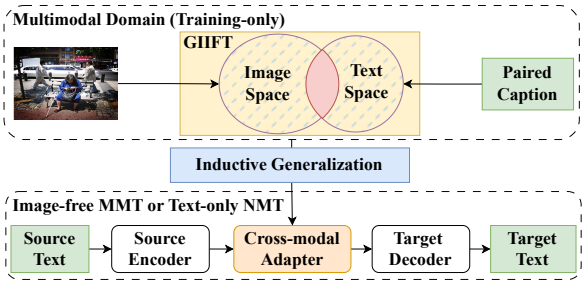


Figure 1: The inductive image-free generalization of GIIFT. GIIFT inductively learns from an entire multimodal domain, then enables image-free inference for MMT or text-only NMT via cross-modal generalization. In contrast, previous models can only learn from the limited overlap between image and text in red.

To address the bottleneck mentioned above, there have been a few methods that attempt to explore resolutions for achieving image-free inference in MMT models. For example, synthesizing visual hallucination from text or textual scene graph during training is used for the image-free inference (Li et al., 2022; Fei et al., 2023); transfer learning from the image-to-text captioning task to the text-to-text translation task (Gupta et al., 2023). However, none of these efforts has managed to consistently reach the performance of fully bridging the gap between multimodal and image-free inference. There are still critical challenges in advancing image-free inference in MMT.

First, previous models learn inadequately because they forcibly align the modality gap, typically the intrinsic information imbalance between im-

[1]Code is available at: https://github.com/xjiaf/GIIFT

Proceedings of the Tenth Conference on Machine Translation, pages 98–112
November 8-9, 2025

ages and texts (Schrodi et al., 2025). Consider the case in Figure 1, this constraint limits image-free inference generalized from only the red overlap in the multimodal domain. Recent work (Ramasinghe et al., 2024; Schrodi et al., 2025), however, shows that accepting this gap and exploiting the full information (the entire collection in Figure 1) substantially enhances multimodal learning. Second, current image-free MMT approaches fail to realize cross-modal generalization, resulting in a marked drop in image-free inference compared with traditional MMT with multimodal inference (Fei et al., 2023). Third, current MMT models are transductive (Sutskever et al., 2014), which struggle with inductively generalizing from multimodal domains to text-only domains for broader applications. Facing these challenges, we investigate the following research questions in this work: *RQ1: How can we fully represent different modalities to embrace the modality gap? RQ2: Can we effectively make image-free inference via cross-modal generalization without downgrading performance? RQ3: Can MMT have the inductive ability to generalize multimodal domains to broader text-only domains?*

We propose **GIIFT**, a two-stage **G**raph-guided **I**nductive **I**mage-**F**ree MM**T** framework. As shown in Figure 1, the GIIFT learns from the entire multimodal domain in the first stage, and aims to achieve inductive generalization for image-free MMT or text-only NMT in the second stage. Graph-structured novel Multimodal Scene Graph (MSG) and Linguistic Scene Graph (LSG) are proposed to represent the multimodal domain, in which each modality informs and enriches the other by graph representation in the unified space. Specifically, we extract visual relationships from images and linguistic relationships from text, then preserve and integrate them via global super nodes to construct MSGs. LSG is a linguistic version, preserving linguistic relationships with global super nodes. These relations' and nodes' features are mutually enriched and uniformly initialized via M-CLIP (Carlsson et al., 2022). To enable cross-modal generalization to image-free or broader text-only domains, GIIFT is designed with a cross-modal GAT adapter to inductively learn multimodal knowledge from the multimodal domain via MSGs in the first stage, and generalize it for image-free inference via LSGs in the second stage, based on a replaceable backbone, mBART (Liu et al., 2020).

Overall, the main contributions are:

(i) We build novel MSGs and LSGs to fully represent different modalities in a unified space for embracing the modality gap.

(ii) We propose the two-stage GIIFT framework with a novel lightweight GAT adapter to achieve inductive cross-modal generalization for image-free inference MMT via MSGs and LSGs.

(iii) Experimental results on En→{Fr, De, Uk} Multi30K show that GIIFT outperforms most of the existing MMT methods even without image during inference. On En→{Fr, De} WMT, GIIFT surpasses the best image-free baseline, CLIPTrans, by average gains of **+1.92 (8.00%) BLEU** and **+2.82 (4.80%) METEOR**, demonstrating effective induction from multimodal Multi30K to other text-only NMT domains.

(iv) Further analysis demonstrates that the proposed GIIFT can effectively embrace modality gaps via MSGs and achieve robust image-free inference via two-stage cross-modal generalization, and shows that the GIIFT can achieve consistent performance of fully bridging the gap between multimodal inference and image-free inference.

2 Related Work

2.1 Multimodal Machine Translation

MMT research integrates visual and textual information for machine translation with an increasing number of models (Grönroos et al., 2018; Huang et al., 2020; Zhao et al., 2022a; Cheng et al., 2024). Early approaches commonly adopted RNN-based architectures enhanced with attention mechanisms to incorporate global or spatial visual features (Calixto et al., 2017). Transformer variants soon supplanted RNNs, introducing tighter cross-modal fusion such as dynamic token re-weighting (Caglayan et al., 2018; Lin et al., 2020), double attention mechanisms (Zhao et al., 2020), gating mechanisms (Wu et al., 2021), multimodal adapters (Zhao and Calapodescu, 2022) or multi-granular fusion via graph structures (Krishna et al., 2017; Wang et al., 2018; Yin et al., 2020). Recent research leverages pre-trained resources such as CLIP (Gupta et al., 2023; Li et al., 2022) or BERT (Li et al., 2020). Within the widely adopted encoder-decoder framework, MMT research has progressed along two fronts in representation and inference:

(1) Visual-linguistic representation. Most MMT models enforce rigid visual-linguistic alignment. Disambiguation work (Ive et al., 2019; Zhao et al., 2022b; Futeral et al., 2023) links each textual token to a matching image region to resolve lexi-

cal ambiguity. UMMT(Fei et al., 2023) aligns every hallucinated visual scene graph node with textual counterpart. CLIPTrans (Gupta et al., 2023) trains sequentially on two stages, image captioning and the corresponding translation, enforcing alignment across both stages. Such alignments discard modality-specific information by merely learning the overlap between modalities.

(2) Image-free inference. Previous MMT relies on access to the paired image at test time. To mitigate this limitation, researchers have pursued three lines of work. First, retrieval-based models replace the missing picture with visual features fetched from an indexed caption-image bank into the decoder (Zhang et al., 2020). Second, hallucination methods (Johnson et al., 2018; Li et al., 2022; Fei et al., 2023) synthesize visual inputs from texts. Third, using transfer learning to train the NMT models with images (Gupta et al., 2023).

2.2 Graph Neural Networks

GNN is powerful in modeling relational structures by leveraging message-passing mechanisms (Wu et al., 2020; Liang et al., 2022), which iteratively aggregates and updates nodes' representation with information from their neighbors, capturing both local and global relational patterns. Some GNNs, such as Graph Convolutional Network (GCN) (Kipf and Welling, 2017) and its variants, are only transductive, while others, including GAT (Veličković et al., 2018), and GraphSAGE (Hamilton et al., 2017), also enable inductive learning (Battaglia et al., 2018; Xiong et al., 2025) to handle previously unseen nodes (Zhou et al., 2020). GNNs also use hierarchical or global pooling techniques (Ying et al., 2018; Lee et al., 2019; Gao and Ji, 2019) to capture subgraph-level or graph-level embeddings (Zhou et al., 2020).

3 Methodology

In this work, we construct unified MSGs and LSGs to generalize multimodal knowledge for image-free inference. We design the GIIFT framework with two-stage continuous learning via a lightweight GAT adapter for inductive cross-modal generalization. Our methodology is structured as follows: 1) We introduce the problem definition of our inductive image-free inference (Subsection 3.1). 2) The details of the MSG and LSG scene graph generation (Subsection 3.2). 3) The description of the GIIFT framework (Subsection 3.3).

3.1 Problem Definition

Let $\mathcal{D}_m$ be a multimodal multilingual dataset of triplets (i, c_s, c_t), where i is an image and (c_s, c_t) are its source and target captions, and let $\mathcal{D}_l$ be a text-only parallel corpus of pairs (t_s, t_t). Traditional MMT methods align i with c_s during training and then perform image-free inference based on that alignment, but the visual knowledge remains tied to c_s and $\mathcal{D}_m$. Our inductive image-free inference instead learns multimodal knowledge from i and (c_s, c_t) in $\mathcal{D}_m$ that cross-modally generalizes to both $\mathcal{D}_m$ or translation pairs $(t_s, t_t) \in \mathcal{D}_l$.

3.2 Scene Graph Generation

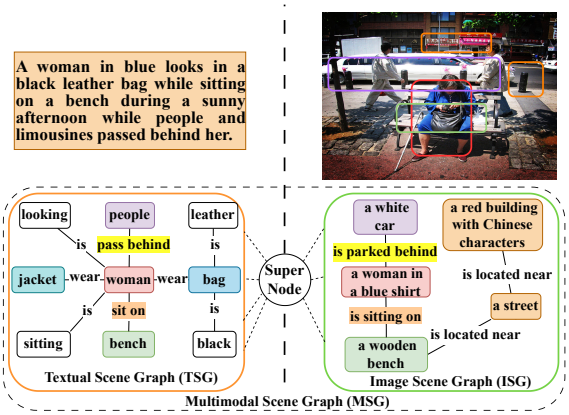


Figure 2: Representation of textual and visual information via MSG. Corresponding objects, attributes, and relations across both Textual Scene Graphs and Image Scene Graphs are depicted using identical colors.

To preserve modality-specific information and extract complex relations (e.g., spatial relations, environmental scene, and event states of people and objects) during data preprocessing, we use a multimodal Large Language Model (LLM) as the image parser and an off-the-shelf text parser to build the MSG and LSG, respectively. Figure 2 shows that the MSG comprises an Image Scene Graph (ISG) and a Textual Scene Graph (TSG), and a visual super node to bridge ISG and TSG included in a unified space, whereas the LSG replaces the visual super node with a textual one and omits the ISG.

(1) Image Scene Graph. ISGs are obtained from images i using LLaVA-34B (Liu et al., 2023). To obtain coherent and well-structured outputs, we set the sampling temperature to 0 for deterministic generations and raise it to 0.4 for more challenging cases requiring exploratory outputs. Our prompts (details in Appendix A) comprise: **(i) Task Description**, specifying how to formulate relations and produce structured scene graph content; **(ii)**

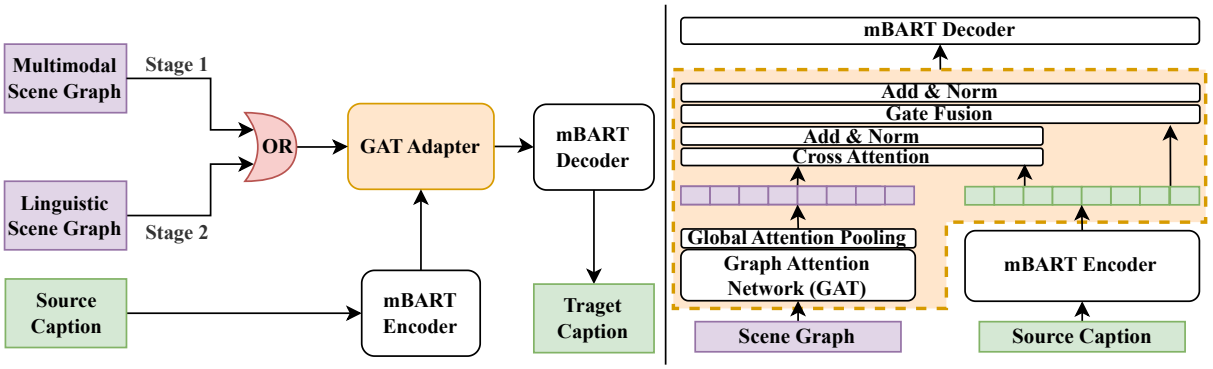


Figure 3: Left: Overview of the two-stage GIIFT framework. Stage 1: multimodal learning via MSGs. Stage 2: cross-modal generalization via LSGs. Right: Overview of the architecture of the cross-modal GAT adapter, which inductively learns and fuses the multimodal knowledge for the backbone, mBART.

Negative Examples, illustrating common errors in output formatting and how to rectify them; **(iii) Format Examples**, providing abstract but well-formed scene graph templates without using specific objects, thus preventing prompt contamination. Therefore, we generate ISGs with unique and relational visual information, such as event states.

(2) Textual Scene Graph. TSGs are parsed by FACTUAL (Li et al., 2023) from texts c_s or t_s. FACTUAL is lighter and more efficient than LLMs for large-scale corpora. TSGs encode entities and their relations as the textual analogue to ISGs. Thus, we obtain TSGs with structured linguistic relationships and unique semantic information.

(3) Multimodal Scene Graph. For each pair (i, c_s), we merge the ISG and TSG into an MSG by introducing a super node that encodes holistic image embeddings from the M-CLIP image encoder. This super node connects to all ordinary ISG and TSG nodes to unite modality-specific information and deliver diverse granular information, serving as a critical bridge to accept the modality gap and build multimodal relationships. We embed all ordinary node and relation features by the M-CLIP text encoder, enabling a unified representation of multimodal relationships and inductive foundations.

(4) Linguistic Scene Graph. For cross-modal generalization, we construct the LSG for text pairs (t_s, t_t) by retaining only the TSG with a super node representing the entire text embedding via the M-CLIP text encoder, which shares the unified hidden space with MSG's. Similarly, we embed all ordinary nodes and edges using the M-CLIP text encoder. The super node connects to all ordinary TSG nodes, enabling multi-granular textual information for image-free inference.

3.3 GIIFT Framework

Harnessing scene graphs as modality-bridges, GIIFT uses MSGs and LSGs, which are inductively learned for multimodal knowledge and generalized for image-free translation, respectively, via an essential cross-modal GAT adapter to guide the backbone mBART in two-stage training pipelines. The proposed lightweight GAT adapter is completely decoupled from the pre-trained mBART backbone (as shown in Figure 3), which can be flexibly incorporated with other language models for deployment.

3.3.1 Two-stage Training Framework

Stage 1: Multimodal Learning via MSG. As shown in Figure 3 (Left), Stage 1 trains a shared GAT adapter on MSGs from paired images and captions, inducing multimodal knowledge that guides image-free translation. Within the MSG, $\mathcal{N}_i^{\mathrm{SG}}$ is the neighbors of the ordinary node i from ISG or TSG, reflecting the structured relationships from image or text. The embedding $\mathbf{Z}_i^{(l)}$ of node i at layer $l, (0 \le l \le L)$ in GAT is calculated recursively as:

$$\begin{aligned}\mathbf{Z}_i^{(l)} = \sigma\Big(\sum_{j\in\mathcal{N}_i^{\mathrm{SG}}} \alpha_{ij}^{(l)}\, \phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)} \| \mathbf{Z}_j^{(l-1)} \| \mathbf{E}_{ij}]) \\ + \alpha_{i,\mathrm{SN}}^{(l)}\, \phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)} \| \mathbf{Z}_{\mathrm{SN}}^{(l-1)} \| \mathbf{E}_{i,\mathrm{SN}}])\Big).\end{aligned} \tag{1}$$

Here, $\mathbf{Z}_i^{(l-1)}$ is the embedding of node i at the previous layer (with initial node embedding $\mathbf{Z}_i^{(0)}$ via M-CLIP text encoder), $\mathbf{Z}_{\mathrm{SN}}^{(l-1)}$ is the global multimodal embedding from the super node at layer $l-1$ which is passed to all the ordinary nodes, and $\mathbf{E}_{ij}$ or $\mathbf{E}_{i,\mathrm{SN}}$ denotes edge embedding of the relationships in the scene graph via M-CLIP text encoder, the $\alpha_{ij}^{(l)}$ and $\alpha_{i,\mathrm{SN}}^{(l)}$ are attention weights, $\sigma(\cdot)$ is

an activate function and $\phi(\cdot)$ is the LeakyReLU. $\mathbf{E}_{ij}$ is the embedding of the relation content in the scene graph, obtained with the M-CLIP text encoder, whereas $\mathbf{E}_{i,\mathrm{SN}}$ is the embedding of the edge between the super-node and each ordinary node i. Because these super-node edges have no textual content, we initialize $\mathbf{E}_{i,\mathrm{SN}}$ as a ones vector whose dimensionality matches that of the embeddings produced by the M-CLIP text encoder.

From Eq. (1), we can observe that the shared weight $\mathbf{W}$ enables learning of multi-granular multimodal relationships by jointly processing local textual embeddings from ISG or TSG nodes ($\mathcal{N}_i^{\mathrm{SG}}$) and global multimodal context from the super node. This shared weight $\mathbf{W}$ is crucial in capturing these multimodal relationships and will be leveraged for the cross-modal generalization process in Stage 2.

The initial MSG super node provides the global image embedding $\mathbf{Z}_{\mathrm{SN}}^{\mathrm{MSG}(0)}$ and aggregates information from all ordinary nodes as follows:

$$\mathbf{Z}_{\mathrm{SN}}^{\mathrm{MSG}(l)}=\sigma(\sum_{i\in\mathcal{N}_{\mathrm{SN}}^{\mathrm{MSG}}}\alpha_{\mathrm{SN},i}^{(l)}\phi(\mathbf{W}[\mathbf{Z}_{\mathrm{SN}}^{(l-1)}\|\mathbf{Z}_i^{(l-1)}\|\mathbf{E}_{\mathrm{SN},i}])) \tag{2}$$

where $\mathcal{N}_{\mathrm{SN}}^{\mathrm{MSG}}$ contains all ordinary nodes in MSG.

Through the Global Attention Pooling (Li et al., 2015) layer in the GAT adapter, we then obtain the multimodal graph representation $\mathbf{Z}_g^{\mathrm{MSG}} \in \mathcal{D}_m$:

$$\mathbf{Z}_g^{\mathrm{MSG}} = \mathrm{AttnPool}(\{\mathbf{Z}_i^{\mathrm{MSG}} : i \in \mathcal{V}_{\mathrm{MSG}}\}), \tag{3}$$

where $\mathcal{V}_{\mathrm{MSG}}$ denotes the set of nodes in the MSG, including ordinary nodes and the super node.

$\mathbf{Z}_g^{\mathrm{MSG}}$ is then fed into the decoder for translation. The encoder remains frozen and the decoder learns a balanced representation to generate translations based on the gate mechanism (see Eq. (8)) between multimodal representations $\mathbf{Z}_g^{\mathrm{MSG}}$ and the source embedding $\mathbf{H}$.

Stage 2: Cross-modal Generalization via LSG. As shown in Figure 3 (Left), Stage 2 inputs the same GAT adapter with LSGs built from texts, allowing multimodal knowledge to be cross-modally generalized to broader image-free domains. In the LSG, each ordinary node i represents a textual entity with the initial embedding $\mathbf{Z}_i^{(0)}$ and a super node of the global text embedding, $\mathbf{Z}_{\mathrm{SN}}^{\mathrm{LSG}(0)}$ by the M-CLIP text encoder. The embedding $\mathbf{Z}_i^{(l)}$ of node i at the layer l for an ordinary node is:

$$\begin{aligned}\mathbf{Z}_i^{(l)} = \sigma(&\sum_{j\in\mathcal{N}_i^{\mathrm{LSG}}}\alpha_{ij}^{(l)}\phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)}\|\mathbf{Z}_j^{(l-1)}\|\mathbf{E}_{ij}]) \\ &+\alpha_{i,\mathrm{SN}}^{(l)}\phi(\mathbf{W}[\mathbf{Z}_i^{(l-1)}\|\mathbf{Z}_{\mathrm{SN}}^{\mathrm{LSG}(l-1)}\|\mathbf{E}_{i,\mathrm{SN}}])).\end{aligned} \tag{4}$$

The LSG super node is updated as:

$$\mathbf{Z}_{\mathrm{SN}}^{\mathrm{LSG}(l)}=\sigma(\sum_{i\in\mathcal{N}_{\mathrm{SN}}^{\mathrm{LSG}}}\alpha_{\mathrm{SN},i}^{(l)}\phi(\mathbf{W}[\mathbf{Z}_{\mathrm{SN}}^{(l-1)}\|\mathbf{Z}_i^{(l-1)}\|\mathbf{E}_{\mathrm{SN},i}])), \tag{5}$$

with $\mathcal{N}_{\mathrm{SN}}^{\mathrm{LSG}}$ denoting all LSG ordinary nodes. The LSGs help the generalization of shared multimodal knowledge weight $\mathbf{W}$ from $\mathcal{D}_m$ in Stage 1 to the image-free domain $\mathcal{D}_l$ in Stage 2.

Similar to Eq. (3), we obtain the graph representation of LSG $\mathbf{Z}_g^{\mathrm{LSG}}$. The mBART decoder is enhanced by generalized knowledge $\mathbf{Z}_g^{\mathrm{LSG}}$ from $\mathcal{D}_m$ and adapted to the image-free domain $\mathcal{D}_l$ with unfrozen mBART encoder hidden state.

3.3.2 Cross-modal GAT Adapter

We employ a multi-layer GAT with residual connections (He et al., 2015; Veličković et al., 2018; Li et al., 2019) to learn multimodal knowledge and cross-modally generalize it to a broader image-free domain. Figure 3 (Right) shows the architecture of the GAT adapter fusing the output from the mBART encoder and enhancing the mBART decoder. We denote $\mathbf{Z}_g$ as the graph representation of MSG or LSG by the Global Attention Pooling. The fusion output $\mathbf{O}$ between the graph representation $\mathbf{Z}_g$ and mBART encoder hidden states $\mathbf{H}$ is performed via the cross attention as:

$$\mathbf{A} = \mathrm{MultiHeadAttn}(\mathbf{H}, \mathbf{Z}_g, \mathbf{Z}_g) + \mathbf{H} \tag{6}$$

$$\mathbf{O} = \mathrm{LayerNorm}(\mathrm{Dropout}(\mathbf{A})) \tag{7}$$

A gate mechanism $\mathbf{g}$ balances the embedding flow:

$$\mathbf{g} = \sigma(\mathbf{W}_2\mathrm{ReLU}(\mathbf{W}_1[\mathbf{O}\|\mathbf{H}])) \tag{8}$$

$$\mathbf{H}' = \mathrm{LayerNorm}(\mathbf{g}\odot\mathbf{O} + (1-\mathbf{g})\odot\mathbf{H}) \tag{9}$$

where $\odot$ denotes element-wise multiplication and $[\cdot\|\cdot]$ represents concatenation, $\mathbf{H}'$ is the input of mBART decoder. This two-stage process demonstrates the central inductive role of the cross-modal GAT adapter in representing and generalizing structured multimodal relationships.

4 Experiments

Datasets. We conduct experiments on two benchmarks: Multi30K (Elliott et al., 2016) and WMT2014 (Bojar et al., 2014). Multi30K is a widely used MMT benchmark as a multilingual extension of the Flickr30k. WMT is a text-only multilingual NMT dataset. For evaluation, we conduct EN→{DE, FR} translation tasks on Multi30K's

Model	EN → DE			EN → FR			Mean Δ
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
mBART (NMT backbone) (Liu et al., 2020)	41.12	36.63	32.89	63.37	57.01	47.28	-2.08
MMT Model with Multimodal Inference							
DCCN (Lin et al., 2020)	39.70	31.00	26.70	61.20	54.30	45.40	-5.41
GMNMT (Yin et al., 2020)	39.80	32.20	28.70	60.90	53.90	-	-5.14
Gated Fusion* (Wu et al., 2021)	42.00	33.60	29.00	61.70	54.80	44.90	-4.13
WRA-MNMT (Zhao et al., 2022b)	39.30	32.30	28.50	61.70	54.10	43.40	-5.02
UMMT# (Fei et al., 2023)	37.40	-	-	56.90	-	-	-7.68
Soul-Mix (Cheng et al., 2024)	**44.24**	37.14	34.26	64.75	57.47	49.25	-0.61
GIIFT (with image)	43.32	37.47	34.66	65.17	**59.11**	**49.76**	-0.21
MMT Model with Image-free Inference							
ImagiT (Long et al., 2021)	38.50	32.10	28.70	59.70	52.40	45.30	-5.68
VALHALLA (Li et al., 2022)	41.90	34.00	30.30	62.20	55.10	45.70	-3.58
VALHALLA* (Li et al., 2022)	42.70	35.10	30.70	63.10	56.00	46.50	-2.78
UMMT (Fei et al., 2023)	32.00	-	-	50.60	-	-	-13.53
CLIPTrans (Gupta et al., 2023)	43.87	37.22	34.49	64.55	57.59	48.83	-0.7
GIIFT (image-free)	44.04	**38.41**	**34.94**	**65.61**	58.05	49.72	

Table 1: BLEU on the Multi30K. Δ is gap vs. "GIIFT (image-free)". "GIIFT (with image)" is trained and tested with images and texts in only one stage with unfrozen mBART. * represent ensembled models, # denotes the model trained and tested with images and texts.

Model	EN → DE			EN → FR			Mean Δ
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO	
mBART (NMT backbone)	69.59	65.07	60.15	82.40	77.63	71.58	-1.35
MMT Model with Multimodal Inference							
DCCN (Lin et al., 2020)	56.80	49.90	45.70	76.40	70.30	65.00	-11.73
GMNMT (Yin et al., 2020)	57.60	51.90	47.60	74.90	69.30	-	-9.72
Gated Fusion* (Wu et al., 2021)	67.80	61.90	56.10	81.00	76.30	70.50	-3.48
WRA-MNMT (Zhao et al., 2022b)	58.30	52.80	48.50	76.30	70.60	63.80	-8.92
UMMT# (Fei et al., 2023)	57.20	-	-	70.70	-	-	-13.42
Soul-Mix (Cheng et al., 2024)	69.93	63.59	59.94	83.24	78.23	73.48	-1.01
GIIFT (with image)	70.65	65.59	61.37	83.32	**78.95**	**73.98**	-0.10
MMT Model with Image-free Inference							
ImagiT (Long et al., 2021)	55.70	52.40	48.80	74.00	68.30	65.00	-11.72
VALHALLA (Li et al., 2022)	68.80	62.50	57.00	81.40	76.40	70.90	-2.92
VALHALLA* (Li et al., 2022)	69.30	62.80	57.50	81.80	77.10	71.40	-2.43
UMMT (Fei et al., 2023)	52.30	-	-	64.70	-	-	-18.87
CLIPTrans (Gupta et al., 2023)	70.22	65.43	61.26	82.48	77.82	72.78	-0.75
GIIFT (image-free)	**71.08**	**65.88**	**61.66**	**83.65**	78.36	73.86	

Table 2: METEOR on the Multi30K. Δ is gap vs. "GIIFT (image-free)". "GIIFT (with image)" is trained and tested with images and texts in only one stage with unfrozen mBART. * represent ensembled models, # denotes the model trained and tested with images and texts.

Model	EN → UK					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
mBART	53.43	74.67	46.05	68.86	44.05	66.75
CLIPTrans	54.01	74.76	46.30	69.23	44.11	66.87
GIIFT (image-free)	**55.10**	**75.18**	**47.85**	**70.01**	**44.93**	**67.05**

Table 3: BLEU and METEOR on EN → UK (Ukraine) task of the Multi30K.

Model	EN → DE		EN → FR	
	BLEU	METEOR	BLEU	METEOR
mBART	15.58	41.18	26.50	52.06
CLIPTrans	16.63	42.13	26.78	51.76
(*w/o.* Stage 1)	17.60	42.81	27.71	53.38
GIIFT (image-free)	**18.10**	**43.88**	**28.70**	**54.58**
(*w/o.* Stage 1)	17.79	43.01	27.89	53.45

Table 4: Comparison of domain generalization on text-only WMT. "(*w/o.* Stage 1)" denotes model trained without image involvement from Multi30K.

three standard test splits: Test2016, Test2017, and MSCOCO. We also train and test EN→{DE, FR} on WMT with multimodal knowledge from Multi30K to evaluate the inductive image-free inference. We downsample WMT train set matching Multi30K's size, while keeping the validation and test sets unchanged.

Implementation Details. We train GIIFT on an A100 GPU, with AdamW optimizer (polynomial decay). The batch size is 64, learning rate is $2e^{-5}$ (Stage 1) and $1e^{-5}$ (Stage 2). The GAT Adapter is 9 layers with the same 1024 dimensions as M-CLIP. Text decoding is beam search with size 5. Implementations are in PyTorch and Huggingface

Transformers library. We report BLEU (Papineni et al., 2001) and METEOR via SacreBLEU (Post, 2018) and the evaluate library (Banerjee and Lavie, 2005), respectively. Results are three-run averages with early-stop patience 5 on BLEU. We chose mBART as our backbone to ensure a fair comparison with baselines such as CLIPTrans (Gupta et al., 2023) and to highlight our improvements. All tables bold the best and underline the second-best. Baseline figures are derived from papers or repositories. All the numbers of each model reported are from their fine-tuned version on the corresponding dataset.

4.1 Results on Multi30K

Table 1 and 2 contain translation performance comparison of BLEU and METEOR on EN→{DE, FR} tasks of Multi30K. Table 3 shows the BLEU and METEOR on EN → UK (Ukraine) task of Multi30K compared with baselines. "GIIFT (with image)" is a single-stage variant where both training and testing use paired images and texts, and the mBART backbone remains fully unfrozen throughout. "GIIFT (image-free)" is the full image-free model trained in two-stage framework as Subsection 3.3.

From Table 1 and 2, we observe that GIIFT (image-free) surpass the strongest baseline, Soul-Mix (even tested with images), with an average gain of +0.61 (1.37%) BLEU and +1.01 (1.55%) METEOR, and over the best scene-graph-based image-free baseline, UMMT, by +13.525 (42.27%) BLEU and +18.865 (36.07%) METEOR. It shows the effectiveness of GIIFT by preserving the entire information from images and texts.

Moreover, the GIIFT (image-free) model with cross-modal generalizing multimodal knowledge for image-free inference and GIIFT (with image) with multimodal inference secure top-two ranks on 5 out of 6 benchmarks with overall parity. GIIFT (image-free) even surpasses GIIFT (with image) by +0.21 (0.63%) BLEU and +0.1 (0.17%) METEOR on average. In contrast, UMMT degrades dramatically without images, scoring on average -5.4 (-12.76%) BLEU and -5.45 (-8.53%) METEOR below UMMT$^{\#}$ with multimodal inference. This demonstrates that the GIIFT has the robustness of cross-modal generalization for image-free inference, which can mitigate the gap between multimodal inference and image-free inference.

From Table 3, we can find that GIIFT (image-free) consistently outperforms both mBART and CLIPTrans baselines on the EN→{UK} task of Multi30K. We can observe that GIIFT achieves an average of +1.45 higher than mBART and +1.15 higher than CLIPTrans in BLEU; and an average of +0.65 higher than mBART and +0.46 higher than CLIPTrans in METEOR. It shows the effectiveness of GIIFT by preserving and learning multimodal information from images and texts for cross-modal generalization.

4.2 Results on WMT

In Table 1, CLIPTrans outperforms other image-free inference baselines, so we adopt it as our primary baseline and use its official repository for experiments. Like GIIFT (image-free), this two-stage mBART-based model is trained with Stage 1 on Multi30K and Stage 2 on WMT. The results of CLIPTrans and GIIFT in Table 4 are obtained under the same model parameter setup.

Table 4 shows that GIIFT achieves the highest BLEU and METEOR scores overall, while also significantly outperforming GIIFT (*w/o.* Stage 1) trained without images from Multi30K on both metrics. This validates GIIFT's inductive ability to achieve robust image-free inference via cross-modal generalization.

The difference between CLIPTrans and GIIFT stems from how each model handles the modality gap and the resulting impact on cross-modal generalization. CLIPTrans attempts to align images to textual captions for transferring the Stage 1 visual features into Stage 2 caption translations via a mapping network, which limits the multimodal correlation for cross-modal generalization. By contrast, GIIFT can effectively embrace the modality gap via graph-guided fusion and achieve inductive generalization via a cross-modal GAT adapter. In Stage 1, all the modalities are learned and represented in a *unified multimodal knowledge space* via multimodal scene graphs (MSGs). In Stage 2, the proposed cross-modal GAT adapter generalizes that knowledge into the text-only domain via the assistance of linguistic scene graphs (LSGs). Benefit from a two-stage inductive learning via MSGs or LSGs, GIIFT shows better performance in achieving robust image-free inference performance.

4.3 Human Evaluation

Table 5 presents human evaluation on the Multi30K EN→FR test set using a 10-point Likert scale for completeness, ambiguity, and fluency, alongside

Model	BLEU	Completeness ↑	Ambiguity ↓	Fluency ↑
mBART	63.37	7.0	7.3	8.1
CLIPTrans	64.55	8.0	5.5	8.8
GIIFT (image-free)	**65.61**	**9.3**	**4.6**	**9.5**
(*w/o.* Stage 1)	64.91	8.8	5.1	9.0

Table 5: Human evaluation metrics (10-point Likert scale) and BLEU scores on Multi30K EN→FR task.

BLEU scores for reference.

Results in Table 5 show that our GIIFT substantially enhances key translation dimensions that correlate with human satisfaction. Compared with the baselines, GIIFT significantly outperforms in completeness, indicating that the model can utilize more multimodal context information. GIIFT achieves the highest completeness, fluency ratings, and the lowest ambiguity among all baselines, demonstrating practically meaningful gains even when BLEU is already high. These gains highlight that our method delivers practical improvements across dimensions that BLEU alone cannot capture.

4.4 Comparison with Multimodal LLMs

Table 6 shows the experimental comparison of BLEU and METEOR on EN→{DE, FR} tasks of Multi30K with multimodal LLMs.

We use an RTX A6000 GPU to employ LlaVA-7B, LlaVA-34B (Liu et al., 2023) and Llama3-70B (Grattafiori et al., 2024) based on Ollama. We adopt LLaVA's few-shot inference paradigm (Brown et al., 2020) by prompting the model with a small set of (image, source caption, target translation) examples drawn from Multi30K and then asking it to translate a new image's caption into the target language.

From Table 6, we observe that our GIIFT achieves the highest BLEU and METEOR scores on EN→{DE, FR} benchmarks, with significant improvements over mBART and LLMs. These results confirm the effectiveness of our GIIFT model and show that compact, domain-specialized multimodal MMT models can outperform larger general-purpose LLMs on the translation task.

Considering the parameter complexity compared with multimodal LLMs, the GIIFT only introduced the GAT adapter in the framework, which is lightweight and efficient for training. For instance, the mBART backbone has approximately 600M parameters, while our GIIFT framework adds only about 33.2M parameters in total, approximately 27M in the GAT adapter layers and 6.2M in the gated fusion layer. Although LLaVA has far more parameters for supporting broad capabilities, our GIIFT's parameters are wholly devoted to yielding greater efficiency and accuracy for the translation.

4.5 Ablation Study

To further verify the effectiveness of the different components in the GIIFT framework, we showed the experimental results of the ablated versions in Table 7 on EN→{DE, FR} Multi30K.

As shown in Table 7, the GIIFT (image-free) with two-stage continuous learning by the proposed GAT adapter achieves the best performance across all benchmarks. Removing the gating mechanism, GIIFT (*w/o.* gate), results in a more significant reduction in BLEU and METEOR scores, underscoring the critical balancing role of the gating mechanism to fuse multimodal knowledge and mBART information. Additionally, omitting the multimodal learning stage 1, GIIFT (*w/o.* Stage 1), leads to a decrease in performance, highlighting the importance of learning generalizable multimodal knowledge from MSGs. The performance of GIIFT (unfrozen) is significantly lower than GIIFT (image-free), and closer to the backbone model mBART. This underscores freezing the mBART encoder in Stage 1 to maintain its stable embedding. Compared with the backbone mBART, there is a substantial drop in BLEU and METEOR, which demonstrates the effectiveness of the GIIFT framework in learning multimodal knowledge and cross-modal generalization for image-free inference.

5 Case Study

To investigate the advantages of GIIFT, we examine cases in comparing: the GIIFT (image-free), which learns multimodal knowledge from MSGs for cross-modal image-free generalization via LSGs; GIIFT (*w/o.* Stage 1), which only learns linguistic relationships from LSGs; and the text-only mBART.

(i) Environmental scene. In Figure 4 (left), GIIFT (image-free) and GIIFT (*w/o.* Stage 1) both correctly capture the spatial preposition "on" through MSG or LSG, respectively, despite the source English caption omitting the spatial information. But lacking the scene graph, mBART produces imprecise translations without "on", which highlights the functions of LSGs to guide the learning and generalize the spatial relation knowledge. This case shows that LSGs guide GIIFT to better generalize spatial relation information from MSGs' multimodal knowledge in Stage 2 for image-free translation inference. Additionally, as shown in

Model	EN → DE						EN → FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
LlaVA-7B	27.15	58.54	23.70	52.05	19.54	47.77	35.67	65.57	34.79	62.94	35.00	62.87
LlaVA-34B	25.30	58.76	25.16	55.58	22.04	51.53	40.25	69.47	38.95	67.36	39.99	68.82
Llama3-70B	40.49	69.09	36.12	65.06	34.38	60.86	50.95	77.13	49.39	74.54	49.38	73.39
mBART	41.12	69.59	36.63	65.07	32.89	60.15	63.37	82.40	57.01	77.63	47.28	71.58
GIIFT (image-free)	**44.04**	**71.08**	**38.41**	**65.88**	**34.94**	**61.66**	**65.61**	**83.65**	**58.05**	**78.36**	**49.72**	**73.86**

Table 6: Comparison with Multimodal LLM on Multi30K.

Model	EN → DE						EN → FR					
	Test2016		Test2017		MSCOCO		Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
mBART (backbone)	41.12	69.59	36.63	65.07	32.89	60.15	63.37	82.40	57.01	77.63	47.28	71.58
GIIFT (*w/o.* Stage 1)	43.63	70.95	37.76	65.49	34.47	60.81	64.91	83.07	57.71	78.01	48.95	73.20
GIIFT (*w/o.* freezing)	42.84	70.37	37.24	65.35	34.38	60.69	63.74	82.62	56.52	77.23	48.92	72.54
GIIFT (*w/o.* gate)	43.50	70.95	37.96	65.11	33.85	60.21	64.14	82.61	57.58	78.00	48.59	72.90
GIIFT (image-free)	**44.04**	**71.08**	**38.41**	**65.88**	**34.94**	**61.66**	**65.61**	**83.65**	**58.05**	**78.36**	**49.72**	**73.86**

Table 7: Ablation study on Multi30K. "GIIFT (*w/o.* freezing)" has an unfrozen mBART encoder in Stage 1. "GIIFT (*w/o.* Stage 1)" is trained without images. "GIIFT (*w/o.* gate)" is trained in two stages without gate fusion.

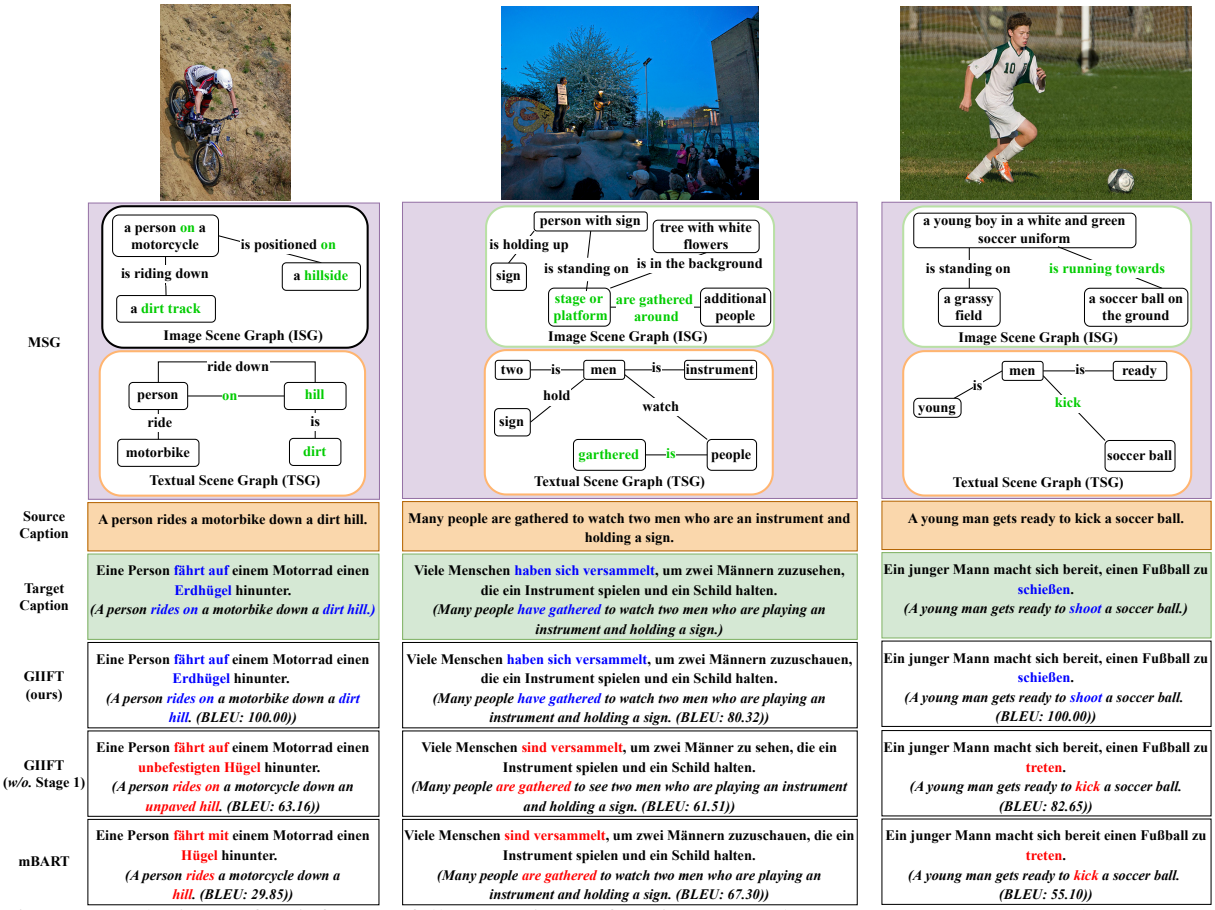


Figure 4: Under image-free inference, full GIIFT (image-free) is compared to GIIFT (*w/o.* Stage 1) and mBART on Multi30K validation set. The italicized bracketed translations of the German caption mark the differences in red.

Figure 4 (left), the environmental scene "dirt hill", is only accurately translated by full GIIFT (image-free) with multimodal knowledge from MSGs. GIIFT (*w/o.* Stage 1) and mBART, although "dirt" in the source caption, produce imprecise translations, reflecting overlooking of the modality-specific information. This case shows the effectiveness of MSGs that can well embrace modality gaps by multimodal graph fusion and fully preserve multimodal information for improving translation. Other cases from Test2016 are shown in Appendix B.

(ii) Temporal states. In Figure 4 (middle), the MSG captures a scene with "are gathered around the stage or platform", enabling full GIIFT to recognize it as a completed state and generate the appropriate perfect tense in German. In contrast,

both GIIFT (*w/o.* Stage 1) and mBART, limited by modality gap, cannot capture the temporal state from an aligned visual-linguistic space, which eliminates the unique temporal state from the image. So they translate English "are gathered" directly into the German present tense.

(iii) Action states. Figure 4 (right) shows the MSG's action state "is running towards" guides GIIFT (image-free) to correctly translate the action as "shoot" rather than "kick". The visual information, available only through MSG, enables the correct translation from multimodal knowledge. GIIFT (*w/o.* Stage 1) and mBART, however, can only literally translate to "kick" in German, which also shows the importance of accepting different modality-specific information rather than alignment.

6 Conclusion

This work introduces GIIFT, a two-stage graph-guided inductive MMT framework, along with novel Multimodal and Linguistic Scene Graphs. GIIFT outperforms existing MMT models on Multi30K even with image-free inference, demonstrating the effectiveness of learning multimodal knowledge in a unified space via MSGs and achieving cross-modal generalization via LSGs. Its image-free performance remains robust, matching its multimodal-inference counterpart. GIIFT also outperforms both the text-only NMT backbone and leading image-free MMT baselines on WMT, showing effective induction of multimodal knowledge to broader text-only domains. Further analysis highlights the advantages of multimodal graph-guided generalization for image-free inference and confirms the effectiveness of the two-stage framework with a lightweight but efficient GAT adapter for cross-modal inductive learning.

Acknowledgments

This work was supported by a Grant-in-Aid for Early-Career Scientists #24K20841, JSPS.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *WMT*, pages 304–323.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, and 8 others. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th international conference on neural information processing systems*, Nips '20, Red Hook, NY, USA. Curran Associates Inc. Number of pages: 25 Place: Vancouver, BC, Canada tex.articleno: 159.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC Submissions for WMT18 Multimodal Translation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and Multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France. European Language Resources Association.

Xuxin Cheng, Ziyu Yao, Yifei Xin, Hao An, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. SoulMix: Enhancing Multimodal Machine Translation

with Manifold Mixup. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11283–11294, Bangkok, Thailand. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *WMT*, pages 215–233.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, Toronto, Canada. Association for Computational Linguistics.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling Ambiguity with Images: Improved Multimodal Machine Translation and Contrastive Evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada. Association for Computational Linguistics.

Hongyang Gao and Shuiwang Ji. 2019. Graph U-Nets. In *International Conference on Machine Learning*, pages 2083–2092.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD Submission to the WMT18 Multimodal Translation Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels. Association for Computational Linguistics.

Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, and Donglai Wei. 2023. CLIPTrans: Transferring Visual Knowledge with Pre-trained Models for Multimodal Machine Translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 1025–1035, Red Hook, NY, USA. Curran Associates Inc. Event-place: Long Beach, California, USA.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling Translations with Visual Awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Justin Johnson, Agrim Gupta, and Li Fei-Fei. 2018. Image Generation from Scene Graphs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1228, Salt Lake City, UT. IEEE.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 123(1):32–73.

Junhyun Lee, Inyeop Lee, and Jaewoo Kang. 2019. Self-Attention Graph Pooling. In *Proceedings of the 36th International Conference on Machine Learning*.

Guohao Li, Matthias Müller, Ali Thabet, and Bernard Ghanem. 2019. DeepGCNs: Can GCNs Go as Deep as CNNs? In *The IEEE International Conference on Computer Vision (ICCV)*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020. What Does BERT with Vision Look At? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, Online. Association for Computational Linguistics.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu (Richard) Chen, Rogerio Feris, David Cox, and Nuno Vasconcelos. 2022. VALHALLA: Visual Hallucination for Machine Translation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated Graph Sequence Neural Networks. *arXiv preprint*. Version Number: 4.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023. FACTUAL: A Benchmark for Faithful and Consistent Textual Scene Graph Parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6377–6390, Toronto, Canada. Association for Computational Linguistics.

Fan Liang, Cheng Qian, Wei Yu, David Griffith, and Nada Golmie. 2022. Survey of Graph Neural Networks and Applications. *Wireless Communications and Mobile Computing*, 2022:1–18.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic Context-guided Capsule Network for Multimodal Machine Translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 1320–1329, New York, NY, USA. Association for Computing Machinery.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742. 01870 Place: Cambridge, MA Publisher: MIT Press.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative Imagination Elevates Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. 2024. Accept the modality gap: An exploration in the hyperbolic space. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27253–27262.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. *Preprint*, arXiv:2404.07983.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *arXiv preprint*.

Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. 2018. Scene Graph Parsing as Dependency Parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 397–407, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Yu Philip S. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24.

Jiafeng Xiong, Ahmad Zareie, and Rizos Sakellariou. 2025. A Survey of Link Prediction in Temporal Networks. *arXiv preprint*.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. 2018. Hierarchical graph representation learning with differentiable pooling. In *Proceedings of the 32nd International Conference on Neural Information Processing*

Systems, NIPS'18, pages 4805–4815, Red Hook, NY, USA. Curran Associates Inc. Event-place: Montréal, Canada.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural Machine Translation with Universal Visual Representation. In *International Conference on Learning Representations*.

Yuting Zhao and Ioan Calapodescu. 2022. Multimodal robustness for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 8505–8516.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *Conference of the European Association for Machine Translation*, pages 105–114.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022a. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022b. Word-region alignment-guided multimodal neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:244–259.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

A Scene Graph Prompts for LLaVA

Please analyze the image provided and construct a structured scene graph, adhering to the following guidelines, and represent it in a JSONL (JSON Lines) format:
1. Entities: List all significant objects or subjects visible in the image, which may include things, animals, or people. Describe each entity in detail, noting their quantities, colors, and any distinctive features. Each description should be distinct and consistent across the document to ensure clarity.
2. Relations: Define all pivotal relationships between the entities using tuples. Each tuple must maintain the exact terminology used in the entities' descriptions. These relationships should be expressed as triplets: [subject entity, predicate, object entity]. Importantly, ensure that the scene graph forms a connected structure. Every entity appearing as a subject or object in one relation must connect to another entity in a different relation, preventing any isolated nodes or subgraphs within the graph. In cases involving an entity related to multiple others, such as being 'between' or 'consist of' them, express this by dividing the relationship into distinct tuples using descriptors like 'is positioned between' and 'and also between' to maintain clarity. Generate triplets with a subject, an active verb or relational word, and a distinct object. Each triplet should clearly describe an action or relationship, avoiding states or implied conditions.
Avoid focusing on too detailed or minor elements that do not significantly contribute to the scene's overall understanding. Use active verbs that show a clear action or relationship. Avoid state or possession verbs like "have" that imply a condition without a distinct action. Incorrect Relations Examples to Avoid:

1.["one person in red shirt", "one dog", "one cat"] (lacks clear action)
2.. . . . . .
Correct Relations Examples of the above, the number of the example is the same as the number of the incorrect example:

1.["one person in red shirt", "is holding", "a book"]
2.. . . . . .
Key Point: Ensure every triplet uses an active verb or distinct relational word to connect the subject and object, clearly describing a specific action or relationship and forming a triplet.

This structure ensures that the scene graph is comprehensive and interconnected, accurately reflecting the dynamics and layout of the scene. The response must strictly follow the JSONL format specified here and not include any extraneous text.

This is a scene graph JSONL example response of the Example Image, the entity_descriptions1, entity_descriptions2, entity_descriptions3, entity_descriptions4 and entity_descriptions5 need to be replaced by specific entities in the image. The relation word1, relation word2, and relation word3 are also need to be replaced by the specific action or relation you observe in the given image. Also, the number of entities and relations is not fixed. It should depend on the given image. The following scene graph JSONL is just an example. You need to describe the real relations based on your given image.

```
{"entities": ["entity_descriptions1", "entity_descriptions2", "entity_descriptions3",
"entity_descriptions4", entity_descriptions5], "relations": [["entity_descriptions1",
"relation word1", "entity_descriptions3"], ["entity_descriptions2", "relation word2",
"entity_descriptions4"], ["entity_descriptions1", "relation word3", "entity_descriptions5"]]}
```

You must not include the word 'image' in the scene graph JSONL. You must not copy the example above! You must describe the entities and their relationships in the given image. Now, you must respond to the scene graph based on the image provided! Straitly follow my instructions. Now what is the scene graph of the image?

We use an RTX A6000 GPU to employ Llava-34B. For each query, the system info explains the task description and provides guidelines for scene graph generation in JSONL format. We also have a quality assessment script to discard any ISG data that fails to meet our generation task description presented in the prompts. We take a temperature of 0 as the default for multimodal large language models (MLLMs) to have a relatively robust performance. If the MLLM can't generate scene graphs that meet the requirements with a temperature of 0, we'll switch to a temperature of 0.4. We exclusively use MLLM for image preprocessing to generate scene graphs according to our task description, with quality ensured by the script.

Source Caption	**Several women are performing a dance in front of a building**	**A crowd gathered around a park water fountain in the rain.**
Target Caption	**Mehrere Frauen führen vor einem Gebäude einen Tanz auf.** ***(Several women perform a dance in front of a building.)***	**Eine Menschenmenge hat sich im Regen um einen Springbrunnen im Park versammelt.** ***(A crowd has gathered in the rain around a fountain in the park.)***
GIIFT (ours)	**Mehrere Frauen führen vor einem Gebäude einen Tanz auf.** ***(Several women perform a dance in front of a building. (BLEU: 100.00))***	**Eine Menschenmenge hat sich im Regen um einen Springbrunnen im Park versammelt.** ***(A crowd has gathered in the rain around a fountain in the park. (BLEU: 100.00))***
GIIFT (*w/o.* Stage 1)	**Mehrere Frauen führen vor einem Gebäude einen Tanz vor.** ***(Several women present a dance in front of a building. (BLEU: 78.25))***	**Eine Menschenmenge versammelt sich im Regen um einen Springbrunnen im Park.** ***(A crowd is gathering in the rain around a fountain in the park. (BLEU: 64.56))***
mBART	**Mehrere Frauen führen einen Tanz vor einem Gebäude auf.** ***(Several women perform a dance in-front-of a building up. (BLEU: 33.03))***	**Eine Menschenmenge hat sich um einen Springbrunnen im Regen versammelt.** ***(A crowd has gathered around a fountain in the rain ~~in the park~~. (BLEU: 45.52))***

Figure 5: Case study of GIIFT on image-free inference when compared to GIIFT (*w/o.* Stage 1) and the mBART. Data points are drawn from the Test2016 set of Multi30K. The gold sentence represents the ground truth. The italicised sentence in the bracket presents the English translation of the German text, while red words highlight the crucial translation differences.

B Case Study on Multi30K Test2016 Set

As shown in Figure 5 (Left), our full model, GIIFT (image-free), correctly associates the text with visual scene context to translate the separable verb accurately, using "auf" to express the "perform" action, whilst GIIFT (*w/o.* Stage 1) incorrectly translates it as "vor" to express "present". mBART fails to properly understand the scene, resulting in disordered word arrangement.

As shown in Figure 5 (Right), our full model, GIIFT (image-free), correctly generalizes the tense from textual to visual information, accurately translating the crowd's gathered state as "has gathered" rather than directly translating "gathered". GIIFT (*w/o.* Stage 1) incorrectly translates it as the progressive tense "is gathered". Although mBART uses the correct tense, it still fails to properly understand the scene context, omitting the location "in the park" and misattributing the modifier "in the rain", leading to overall semantic confusion.

Specification-Aware Machine Translation and Evaluation for Purpose Alignment

Yoko Kayano[1,2] **Saku Sugawara**[1,2]
[1]The Graduate University for Advanced Studies (SOKENDAI)
[2]National Institute of Informatics
{yokokayano,saku}@nii.ac.jp

Abstract

In professional settings, translation is guided by communicative goals and client needs, often formalized as specifications. While existing evaluation frameworks acknowledge the importance of such specifications, these specifications are often treated only implicitly in machine translation (MT) research. Drawing on translation studies, we provide a theoretical rationale for why specifications matter in professional translation, as well as a practical guide to implementing specification-aware MT and evaluation. Building on this foundation, we apply our framework to the translation of investor relations texts from 33 publicly listed companies. In our experiment, we compare five translation types, including official human translations and prompt-based outputs from large language models (LLMs), using expert error analysis, user preference rankings, and an automatic metric. The results show that LLM translations guided by specifications consistently outperformed official human translations in human evaluations, highlighting a gap between perceived and expected quality. These findings demonstrate that integrating specifications into MT workflows, with human oversight, can improve translation quality in ways aligned with professional practice.

1 Introduction

High-quality translation in professional settings requires more than a literal rendering of the source text. It must also fulfill a communicative purpose, which depends on factors such as the intended function, target audience, and the broader context of the original text (Reiss and Vermeer, 1984; Nord, 2006). A single source text may yield different translations depending on these factors.

These contextual factors are typically documented as *translation specifications*. A specification is a predefined set of conditions that guide the translation process, including purpose, audience, tone, style, and content priorities (ISO17100:2015; JTF, 2018). They help translators make informed decisions and ensure that the translation meets user needs (Reiss and Vermeer, 1984; Nord, 2006). Without such guidance, translators may struggle to begin the process at all.

Specifications are also essential in translation evaluation. Frameworks such as ISO 5060 and the Multidimensional Quality Metrics (MQM) emphasize specification-based assessment (ISO5060:2024; Lommel et al., 2013). Lommel et al. (2013) state that "translation quality can only be assessed in terms of whether or not a translation meets specified requirements and meets its communicative purpose." When specifications are absent or vague, evaluations tend to focus on surface-level features such as lexical accuracy or fluency, rather than on whether the translation achieves its communicative purpose. This perspective is also central to functionalist theories in translation studies, which hold that quality should be judged by how well a translation fulfills its intended purpose in the target context, rather than by equivalence with the source text (Reiss and Vermeer, 1984).

MQM is widely adopted in machine translation (MT) research, including the Conference on Machine Translation (WMT), where it underpins human evaluation (Freitag et al., 2021, 2024; Zerva et al., 2024). Its detailed error typology has contributed to translation evaluation. However, specifications are often treated implicitly, and the idea of translation as a goal-oriented process is not fully integrated into MT research. As a result, MT outputs often fall short in real-world applications where purpose and audience matter. This gap is increasingly problematic as industry clients now expect translations to serve specific business objectives (Lommel et al., 2024a).[1]

In response, we propose a framework for specification-aware MT and evaluation. Figure 1 outlines this framework, contrasting traditional

[1]See Appendix A for further discussion.

Proceedings of the Tenth Conference on Machine Translation, pages 113–141
November 8-9, 2025

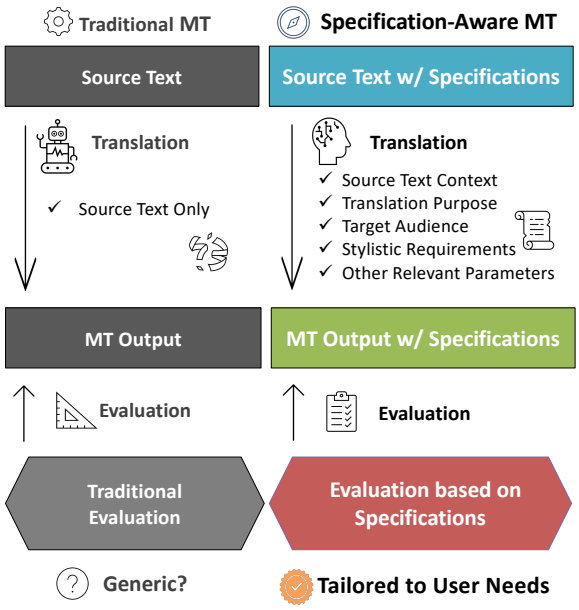


Figure 1: Comparison of traditional MT pipelines, based solely on the source text, and specification-aware pipelines that incorporate contextual and purpose-specific information.

MT pipelines with our approach, which incorporates contextual information and specification-based evaluation. This approach reflects the functionalist view that translation should be guided by purpose, audience, and context. Incorporating specifications into MT workflows and evaluation is essential for improving translation quality in professional domains. To support this claim, we present the theoretical foundation for understanding the role of specifications in Section 3. We introduce a practical guide based on international standards such as ISO 5060, ISO 17100, MQM, and the JTF guidelines in Section 4 (ISO5060:2024; ISO17100:2015; MQM, 2025; JTF, 2018). We apply this framework to a case study in Section 5.

Our case study focuses on investor relations (IR) texts from 33 publicly listed Japanese companies, where alignment with specifications is crucial. We compare five English translations: the company's official version, a proprietary MT output, a basic prompt-based LLM output, a prompt-based LLM output using specifications, and a prompt-based LLM post-edit of the MT output.[2]

We evaluate the translations through expert error analysis, user rankings, and a reference-free automatic metric. Results show that prompt-based LLM outputs with specifications receive the highest ratings from both experts and users. Official translations score lower due to stylistic shortcomings, and conventional MT outputs also underperform. In contrast, the automatic metric favors MT output, highlighting a misalignment with human judgment. This suggests that specification-aware LLM outputs better fulfill communicative goals, even if not fully reflected in current automatic metrics.[3]

Our contributions are as follows:

- We provide a theoretical foundation for incorporating translation specifications into MT and its evaluation.
- We propose a practical guide for applying specifications in MT workflows and test it on IR texts.
- We demonstrate that specification-aware prompt-based LLMs outperform official human translations in human evaluations, supported by detailed analysis.

2 Related Work

2.1 Customizable Machine Translation

Recent studies have explored how MT can be adapted to specific contexts using external knowledge, prompts, and post-editing beyond the source text.[4] Though not framed as translation specifications, these efforts share similar goals with ours.

Fujita (2021) highlights the limits of text-to-text neural machine translations (NMT), emphasizing the need for style guides, terminology, and domain knowledge. He (2024) and Jiao et al. (2024) show that prompting GPT-4 with contextual cues and post-editing improves translation quality. Liu et al. (2025) further finds that detailed, domain-specific prompts enhance performance in specialized tasks.

For stylistic and functional control, Moslem et al. (2023), Wang et al. (2023), and Yamada (2023) show that incorporating tone, terminology, and information about the translation's purpose and audience leads to more targeted outputs. Raunak et al. (2023) also demonstrate that post-editing

[2]Throughout this paper, *LLM* refers to large language models guided by prompt-based customization. We use *post-edit* in a broad sense, including automated revision, and not limited to human editing as defined in ISO 18587 (ISO 18587:2017, 2017).

[3]All translation data and evaluation results are available at https://github.com/nii-cl/Specification_aware_MT.

[4]For a discussion of prior work on controllable MT in the Statistical Machine Translation (SMT) and NMT eras, see Appendix B.

GPT-4 output improves English–Chinese and English–German translation quality.

These studies share motivations with our work on specification-aware translation. Our study extends these efforts by using real corporate materials and evaluating translation outputs in a practical setting. We explore the potential of prompt-based LLMs to meet specific professional translation requirements, based on human evaluation.

2.2 Advances in Translation Evaluation

Recent work in translation evaluation moves beyond gold references and explores reference-free approaches (Blain et al., 2023; Freitag et al., 2023, 2024; Zerva et al., 2024). Evaluation criteria also expand to include contextual coherence and fine-grained error types. For document-level automatic evaluation, Vernikos et al. (2022) improve sentence-level metrics by incorporating context. Jiang et al. (2022) propose BlonDe, a metric that evaluates discourse coherence using span-level F1 scores. Meta-evaluation, such as that of Moghe et al. (2025), examine whether metrics can detect diverse error types and highlight their limitations.

MQM-based automatic evaluation has gained traction: GEMBA-MQM (Kocmi and Federmann, 2023a), AutoMQM (Fernandes et al., 2023), and xCOMET (Guerreiro et al., 2024) identify error spans and types without language-specific tuning. CATER (Iida and Mimura, 2024) offers reference-free, multi-dimensional evaluation with LLMs, while MQM-APE (Lu et al., 2025) adds automatic post-editing to LLM-based error annotation to focus on quality-improving edits.

Human evaluation also remains essential. Lommel et al. (2024b) present an MQM scoring framework with calibrated models. MQM-Chat (Li et al., 2025) adapts MQM for chatbot, and ESA (Kocmi et al., 2024) streamlines span-level annotations for non-expert assessments.

Building on these developments, our study incorporates ISO 5060 and MQM principles into a specification-aware evaluation framework. To better capture translation quality, we assess translations using expert annotators, end-user judgments, and automatic metrics, highlighting both linguistic quality and functional adequacy.[5]

[5]Appendix C provides more context on discussions of evaluation method reliability and improvement in NLP.

2.3 Translation Theory and Machine Translation

Several studies explore interactions between translation studies and MT research. Tan et al. (2023) apply Skopos-based criteria to compare human and NMT outputs, showing that human translations perform better due to NMT's contextual and lexical limitations. Liu et al. (2024) recommend integrating Skopos theory into human evaluation, while Na et al. (2024) show that theory-informed prompts affect LLM outputs. Hiebl and Gromann (2023) call for a unified concept of translation quality to support collaboration between the fields.

The point raised by Hiebl and Gromann (2023) is important: clarifying how the two fields define and evaluate translation quality may help advance both. To this end, we combine theoretical insights from translation studies with empirical experiments based on real-world workflows, aiming to explore how MT can better address the practical needs of professional translation. This integration of theory and practice enables a more realistic understanding of MT's role in professional contexts.

3 Theoretical Background

While translation is often seen as producing an equivalent text in another language, the notion of *equivalence* has faced criticism in translation studies since the late 1970s. In response, functionalist approaches have gained prominence, viewing translation as a purpose-driven communicative act. Skopos theory, a widely cited framework, holds that translations should be guided by their purpose. Based on this view, we argue that translation specifications are essential for developing and evaluating MT systems that meet real-world goals.

We begin with equivalence theory, which frames translation as reproducing the meaning or value of the source text, a view reflected in early MT systems and many current automatic evaluation methods (Section 3.1). We then turn to Skopos theory, a functionalist perspective aligned with our emphasis on translation specifications (Section 3.2). Finally, we discuss how specifications matter not only for translation but also for evaluation (Section 3.3).

3.1 Equivalence Theory in Translation Studies and Machine Translation

Equivalence theory (Nida, 1964), which views translation as reproducing the source text's meaning and value, has long been central to translation

studies.[6] Dyvik (1992) explores this concept in MT, proposing a *situation schema*, an abstract representation that links source and target texts through shared meaning. He emphasizes the importance and difficulty of achieving equivalence, even with linguistic theories and technology.

In contrast, Hardmeier (2015) analyzes how statistical machine translation (SMT) operationalizes equivalence through techniques such as word alignment and domain modeling. While SMT reflects equivalence-based assumptions, he argues it oversimplifies translation complexity.

These studies illustrate how earlier MT systems, primarily rule-based and statistical, were shaped by equivalence-oriented thinking. Although neural networks and deep learning emerged in the mid-2010s (Bahdanau et al., 2015), earlier systems dominated the field and adhered to formal equivalence.

Even with advances in MT technology, equivalence continues to shape quality evaluation. Metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) assess similarity to references, often through n-gram overlap, reflecting a formal equivalence perspective. Recent model-based metrics like COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020) seek semantic equivalence, yet rely on source-text alignment.

Since the late 1970s, however, equivalence theory has faced criticism. Snell-Hornby (1995) argues that it lacks precision and falsely implies symmetry between languages. She identifies the 1980s cultural turn as a shift from language-based approaches to views considering sociocultural context and the translator's role (Snell-Hornby, 2006).[7]

Although equivalence has become less central in translation studies, it is still prominent in MT practice and evaluation. While semantic equivalence remains foundational, it does not fully address the diverse purposes and communicative contexts of real-world translation. Functionalist approaches, such as Skopos theory, offer a useful complement. Emphasizing the intended function, Skopos theory provides a more practical framework for guiding both human and MT in applied settings.

[6]Equivalence theory includes various perspectives, including formal equivalence, which preserves structure, and dynamic equivalence, which aims for a similar reader response (Nida, 1964; Munday et al., 2022; Pym, 2023).

[7]This shift is reflected in major anthologies such as *The Translation Studies Reader* (Venuti, 2021), whose fourth edition retains only Nida (1964) for equivalence theory, omitting figures like Vinay and Darbelnet (1958) and Catford (1965).

3.2 Skopos Theory and the Functional Approach to Translation

Skopos theory defines translation not as a linguistic transfer but as an intentional activity to fulfill communicative goals (Reiss and Vermeer, 1984). Nord (2006) develops this perspective by introducing translation brief (or specification), a set of instructions outlining the purpose, audience, and conditions for the translation. She emphasizes that translation decisions are not solely determined by the source text, but by how well it serves its function.

Gouadec (2007) applies the functionalist approach to translation workflows by identifying three criteria: the client's objectives (e.g., increasing sales or enhancing brand image), the user's needs (e.g., clarity in technical documentation), and the relevant usage norms and standards. As Pym (2023) notes, this positions translators as "language technicians" who operate within a broader communication strategy, ensuring that the translation fulfills its specific role.

This functionalist perspective, once limited to human translation, may now extend to MT with the emergence of prompt-based LLMs. Earlier domain-specific MT systems required significant resources, including specialized datasets, expert tuning, and time-consuming model training (Saunders, 2022; Wang et al., 2023). In contrast, prompt-based LLMs allow users to specify translation requirements through prompting or fine-tuning, making it easier to adapt translations to their intended purpose (Section 2.1). Although empirical evidence is still emerging, customization is now easier, and the rise of LLMs marks a technological shift that aligns with the functionalist view of translation.

Our study investigates whether and how recent advances in MT, especially prompt-based LLMs, can support a functionalist approach by producing translations aligned with specifications. This perspective shifts the focus from linguistic equivalence to functional effectiveness and offers insights for improving MT design and evaluation.

3.3 Why Specifications Matter

Specifications may include parameters such as purpose, target audience, style, register, domain, timeline, cost, volume, reference materials (e.g., glossaries and style guides), file format, and quality evaluation methods (ISO17100:2015; JTF, 2018). In professional settings, such specifications guide translation decisions and ensure the translation

	Parameter	Description
1	Purpose of translation	Communicative goal (e.g., inform, persuade, comply, etc.)
2	Target audience	Intended readers and their language background or expectations
3	Style, register, and tone	Formality, style, and tone appropriate for the target context
4	Terminology and reference resources	Use of glossaries, style guides, and prior translations
5	Domain and legal requirements	Industry norms and compliance with relevant laws
6	Cultural adaptation	Adjustments for cultural norms or sensitivities
7	Length and formatting	Constraints on text length, layout, or structure
8	Localization needs	Regional or language variant customization

Table 1: Translation specification parameters. Items 1–3 are essential; others may vary by project.

meets client expectations. Even if undocumented, essential requirements are typically agreed upon in advance and vary by project. For example, legal translations emphasize consistency with terminology and style guides, while marketing texts prioritize creativity and persuasive language.

The importance is also reflected in how translation quality is evaluated. The MQM framework defines translation quality as follows:

> A quality translation demonstrates required accuracy and fluency for the audience and purpose and complies with all other negotiated specifications, taking into account end-user needs (Melby, 2012a; Lommel et al., 2013).

As explained in the Multi-Range Theory (Lommel et al., 2024b), quality evaluation begins with an analysis of project specifications and user needs. Evaluators should *select* appropriate error categories and scoring models based on this analysis. These ideas are emphasized in the 2024 MQM anniversary paper (Lommel et al., 2024b).[8]

Specifications not only guide translators but also constrain the range of acceptable choices, helping to reduce subjectivity. As Gouadec (2007) argues, translators are language technicians whose "plurality is his enemy," highlighting the importance of clear instructions (Pym, 2023). This applies equally to evaluation: assessments grounded in specifications are less influenced by personal interpretation.

Research shows that providing clear criteria and context improves inter-annotator agreement (IAA) (Castilho, 2021; Popović, 2021). The official MQM website also notes that the framework supports standardized, objective evaluation by minimizing subjective judgment (MQM, 2025). For details on how translation specifications can be incorporated into MQM-based evaluation, see Appendix E.

[8]The MQM framework draws on Garvin (1984)'s approach to quality. See Appendix D for further explanation.

4 A Practical Guide for Specification-Aware MT and Evaluation

We provide a brief overview of our practical framework for integrating translation specifications into both MT workflows and evaluation, where MT is performed using prompt-based LLMs.[9] A full version is available in Appendix F.

4.1 Specification-Aware Machine Translation with Prompt-Based LLMs

Step 1: Define Specifications Clarify translation requirements. These form the basis for both machine output and human review. Our specification parameters, listed in Table 1, are independently developed based on professional translation practice and informed by existing standards and research (ISO17100:2015; 11669:2024; Melby, 2012b).

The top three items are essential for all translation projects, regardless of domain or medium. The remaining items are project-dependent and may be included as needed. Additional parameters may be added depending on the context or client requirements. A brief explanation and examples for each item are provided in Appendix F.1.

Step 2: Design Instructions Specifications should be reflected in prompts or fine-tuning. It is important that the instructions also include source text information, target language, and relevant specification parameters, while preventing hallucination and over-generation.

Step 3: Generate and Review Use LLMs to generate the translation, followed by human review to ensure the output meets specifications. Reviewers make corrections and finalize the translation.

[9]We base our translation and evaluation guidelines on a typical professional workflow and ISO 17100 for translation, and on ISO 5060, JTF guidelines, and MQM for evaluation (ISO17100:2015; ISO5060:2024; JTF, 2018; MQM, 2025).

4.2 Specification-Aware Evaluation

Translation evaluation is not always conducted alongside the translation itself. It may be required in various contexts, such as accepting or rejecting a translation, comparing outputs, selecting the best version, ensuring quality in professional workflows, evaluating MT results, or training and certifying translators. We outline a framework that incorporates both objective and subjective evaluations.

Step 1: Make Specifications Accessible Ensure all evaluators have access to the translation specifications. If not provided in advance, define them before evaluation begins.

Step 2: Define Error Categories Set error categories (e.g., Accuracy, Style, Terminology, etc.) aligned with the specifications. Use established frameworks such as MQM and ISO 5060 (MQM, 2025; ISO5060:2024).

Step 3: Weight and Score Errors Assign weights to error categories based on project priorities. Evaluate severity (e.g., minor, major) and calculate a total score using a weighted formula.

Step 4: Add Subjective Evaluation (Optional) In addition to error-based scoring, subjective evaluation helps assess whether a translation is appropriate, persuasive, and effective for its intended audience. Feedback from experts or users can offer insights into clarity, tone, and impact that error metrics alone may overlook.

The following case study demonstrates the application of this practical guide.

5 A Case Study in Japanese-to-English Translation of Investor Relations Materials

We present a case study to show how specification-aware translation can be applied using a prompt-based LLM. We compare it with human translation and non-prompt-based MT outputs, examining how each is evaluated through both human and automatic methods. This case study puts into practice the notions discussed earlier in Section 3 and assesses the effectiveness of our approach in a real-world Japanese-to-English translation task.

Figure 2 provides an overview of the case study. For LLM, we use ChatGPT via its public interface to simulate a scenario in which non-expert users, such as translators or corporate communications personnel, can control translation output through prompting, without needing specialized tools or programming skills.

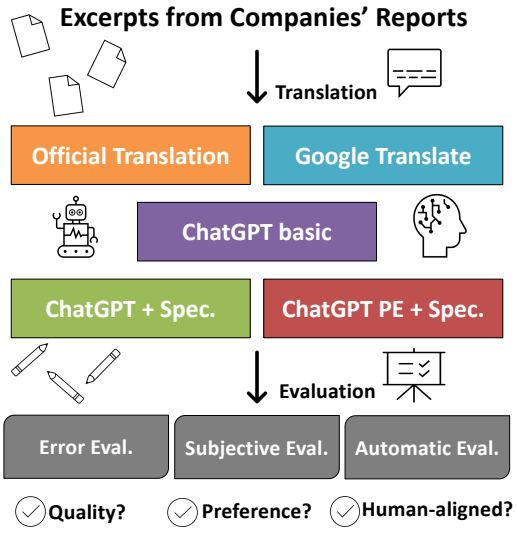


Figure 2: Overview of the case study: Five translation types and their evaluation via expert, user, and automatic methods.

5.1 Experimental Setup

5.1.1 Integrated Reports

We use IR materials excerpted from integrated reports by publicly listed Japanese companies as the source text. The focus is on Japanese-to-English translation; the rationale for using this language pair is explained in Appendix G.

Integrated reports combine financial and non-financial information to communicate a company's value creation to investors and other stakeholders. Although not legally required in Japan, their publication has increased with growing interest in ESG investment. As of the end of 2023, 1,019 companies issue integrated reports, 70 percent of which also provide English versions (ESG/Integrated Reporting Research Laboratory, 2024). Among Prime Market companies, over half publish integrated reports. We choose integrated reports because they are more structured than websites and more interpretive than financial statements, posing challenges for both human translators and MT systems.

We focus on the corporate philosophy section, typically found at the beginning of these reports. According to the *Guidance for Collaborative Value Creation 2.0* (Ministry of Economy, Trade and Industry, 2022), such statements are central to investor communication and must clearly express a company's unique values. Translating them re-

Type	Translation Output
Source	“ワクワク”は, 人を動かすエネルギー. それは人から人へと伝わり、世界をあかるく元気にする。
Official	“Waku waku” is what moves people to push what’s possible. It’s Japanese for the joy and excitement of discovering the unknown. And when passed from person to person, becomes a force that creates a brighter world, united in wonder.
Google	“Excitement” is the energy that moves people. It spreads from person to person, making the world brighter and more energetic.
GPT-b	“Excitement” is the energy that moves people. It spreads from person to person, bringing brightness and vitality to the world.
GPT+Sp	Excitement is the energy that moves people. It spreads from person to person, brightening and invigorating the world.
PE+Sp	“Excitement” is the spark that moves people, spreading from one person to another, brightening and energizing the world.

Table 2: Differences in translations (All Nippon Airways Co., Ltd.)

quires not only literal accuracy but also clarity, an appropriate corporate tone, and expressions that enhance appeal to stakeholders.

We select one company from each of the 33 industries defined by the Tokyo Stock Exchange (Japan Exchange Group, Inc., 2021), prioritizing those with higher market capitalization (27 first-ranked, five second-ranked, and one fourth-ranked).[10] The extracted sections range from 240 to 927 Japanese characters, with an average of 610.

We manually confirmed alignment with the Japanese source texts and asked companies how their English versions were produced. Of the 33 companies, 20 responded. Among these, 15 used only human translation, two combined human and MT, two declined to disclose their method, and one outsourced the work without providing details. None reported using MT alone, suggesting that human translation remains standard.

5.1.2 Five Translation Methods

To compare translation quality and effectiveness, we prepare five versions using different methods. The official translation consists of excerpts from English versions of integrated reports published by the companies.

We then create four MT-based versions:

- **Google Translate**: raw output from Google Translate
- **ChatGPT basic**: ChatGPT with a minimal prompt
- **ChatGPT + Spec**: ChatGPT with specifications
- **ChatGPT PE + Spec**: Google Translate post-edited by ChatGPT with specifications

To ensure consistency, we use the first output for all versions. All ChatGPT translations are generated using ChatGPT-4o.

[10]The full list appears in Appendix H.

For the specification-aware methods, we provide prompts that reflect key information such as source text context, intended purpose (e.g., appealing to global investors), target audience, and stylistic tone. The full prompt is shown in Appendix I.

Using these methods, we generate five translations for each of the 33 companies. Manual review indicates that all versions maintain overall meaning without serious accuracy errors. However, we observe a recurring issue in ChatGPT translations: kanji misinterpretation. For example, 文殊院旨意書 (Monjuin Shiigaki) is rendered incorrectly as *Monjuin Shiisho*, lacking accurate transliteration. This suggests that ChatGPT struggles with domain-specific terminology and proper nouns.

Table 2 shows translations of a corporate philosophy excerpt from All Nippon Airways Co., Ltd.’s integrated report. The official translation contains a grammatical error (“And when passed... becomes...”) and phrases that may be unclear or unnatural (“to push what’s possible”). It also gives an extended explanation of *waku waku*. The Google translation is grammatically correct but closely mirrors the source, resulting in a literal tone and basic vocabulary. The ChatGPT basic version improves fluency and uses slightly richer expressions (“vitality”), but its tone and structure remain similar to the Google version. The ChatGPT version with specifications uses more active verbs and parallel phrasing (“brightening and invigorating”), resulting in smoother rhythm and tone. The post-edited version with specifications introduces vocabulary like “spark” and “energizing,” while preserving the original meaning and structure.

These examples show that each method yields distinct results and that adding specifications to ChatGPT prompts may encourage more purposeful and expressive language. Appendix J provides a comparative analysis of a longer excerpt from the same source, focusing on linguistic and stylistic differences.

5.2 Human Evaluation

After preparing translations for all 33 companies, we conduct two human evaluations of the five translation methods: expert error evaluation and subjective evaluation.

5.2.1 Error Evaluation

We conduct an error-based evaluation using the specification-aware framework introduced in Section 4.2 and detailed further in Appendix F.2. This evaluation focuses on three core categories: Accuracy, Linguistic Conventions, and Style. Other categories defined in the MQM framework, such as Design and Markup, are excluded as they are not applicable to the scope of this study.

All category definitions are based on the MQM standard and were provided to the evaluators to ensure consistency and shared understanding (MQM, 2025). See Appendix K for the detailed error categories used in the annotation.

Given the importance of stylistic quality in IR materials, we include four subtypes under Style: (1) Language register mismatch, (2) Awkward style, (3) Unidiomatic expressions, and (4) Inconsistent style. These errors do not hinder comprehension but result in unnatural English that may reduce clarity and impact. Subtypes help clarify scope, but annotators classify errors only at the main category level to reduce cognitive burden. Error categories are weighted based on JTF guidelines: Accuracy (0.7), Linguistic Conventions (0.8), and Style (1.5), averaging to 1.0 overall (JTF, 2018). We do not apply severity levels, as the texts do not involve high-stakes content such as financial figures.

Two professional evaluators, each either a professional translator or an expert in linguistics and culture, bilingual in Japanese and English, and with English as their first language, are recruited via Prolific.[11] They receive the Japanese source text, translation specifications, five anonymized English translations, an error typology table with definitions, and sample annotations. They identify errors, assign them to one of the three categories, mark their locations, and record error counts. Each evaluator is compensated £40 for approximately 270 minutes of work. Only two out of 24 recruited participants completed the task, highlighting the practical difficulty of securing qualified evaluators and the cognitive demands of error annotation, as noted in prior research (Kocmi et al., 2024; Zouhar et al., 2025).

[11] https://www.prolific.com

Type	Official	Google	GPT-b	GPT+Sp	GPT PE+Sp
Eval. 1	*2.60*	1.82	1.04	0.70	**0.38**
Eval. 2	*3.01*	2.29	1.28	1.29	**1.03**

Table 3: Weighted error scores averaged across 33 companies. Lower scores indicate fewer errors and higher translation quality.

Table 3 presents the evaluation results. ChatGPT PE + Spec receives the lowest error score (highest quality), followed by ChatGPT + Spec, ChatGPT basic, and Google Translate. The official translation ranks lowest, with particularly frequent Style errors, which will be discussed in Section 5.2.3. These findings suggest that LLM-based translations guided by specifications can outperform human translations in this context, challenging the assumption that human translations should serve as the default gold standard in MT evaluation.

We also assessed inter-annotator reliability by calculating the correlation between the error scores assigned by the two evaluators. Pearson's correlation is very high (r = 0.985 and p = 0.0021), while Spearman's rank correlation is also strong (ρ = 0.90 and p = 0.037), indicating statistically significant agreement. Nonetheless, we observe inconsistencies: the same expression was sometimes marked as an error in one translation but not in another by the same evaluator. This indicates the inherent difficulty of ensuring consistency in error-based evaluation, even among professionals. The low completion rate suggests that translation evaluation is time-consuming, cognitively demanding, and difficult to delegate, as it requires a high level of expertise. Our evaluation process incidentally reflected these challenges in practice.

5.2.2 Subjective Evaluation

To understand how translations are perceived by intended end users, we conduct a subjective evaluation alongside expert-based error analysis. As discussed in Section 4.2 and detailed in Appendix F.3, combining error-based and subjective evaluation is useful not only when qualified annotators are limited, but also when end-user perspectives take precedence. For texts like integrated reports, which aim to build trust and attract investment, reader impression may matter more than linguistic accuracy, making subjective feedback particularly valuable.

Subjective evaluation is generally divided into expert and end-user perspectives (JTF, 2018).

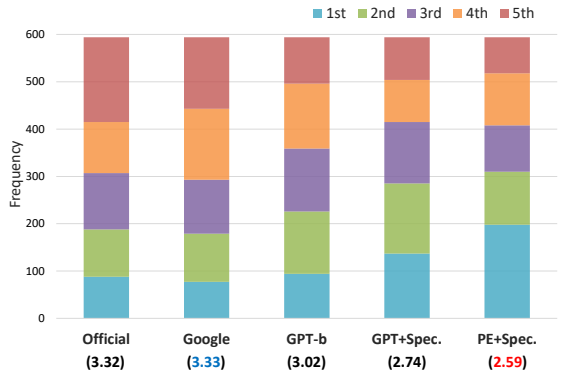


Figure 3: Ranking counts from the subjective evaluation. The x-axis shows translation methods, and the y-axis shows frequency. Each bar is stacked by rank (1st to 5th). Numbers in parentheses indicate mean rankings; lower values reflect higher preference.

Since integrated reports target investors, we adopt an end-user perspective. We recruited eighteen native English speakers via Prolific. Participants are compensated £13.50 for approximately 90 minutes of work. Seventeen hold degrees in fields such as accounting or finance. One participant is a translator and linguistic expert who also took part in the error evaluation. Each evaluator receives the translation specifications and five English translations, without knowing the translation methods or having access to the Japanese source. They are asked to rank the translations based on overall appeal, defined as clarity, readability, word choice, and company presentation.

Figure 3 shows the distribution of rankings across translation types. ChatGPT PE + Spec is most often ranked 1st, whereas the official translation is most often ranked 5th. Google Translate is the least often ranked 1st, while the three ChatGPT-based translations are less often ranked last. Numbers in parentheses indicate mean rankings: ChatGPT PE + Spec has the best (lowest) average ranking, followed by ChatGPT + Spec, ChatGPT basic, and the official translation. Google Translate ranks lowest overall.

To assess significance, we conduct Wilcoxon signed-rank tests on all ten translation pairs, reporting the test statistic (W), p-values, and effect sizes (r) in Table 4. A p-value below 0.05 is considered significant; effect sizes are interpreted as small ($r = 0.1$), medium ($r = 0.3$), or large ($r = 0.5$). All comparisons are significant except Official vs. Google Translate and ChatGPT + Spec vs. ChatGPT PE + Spec. Pairs with ChatGPT PE + Spec (vs. Official and Google Translate) show the strongest effects, with $p \approx 0$ and medium effect sizes.

Pair	W	Z	p	r
Off. vs Ggl	88018	0.081	0.9341	0.003
Off. vs GPT-b	74913	3.213	0.00113	0.132
Off. vs GPT+Sp	**63954**	**5.832**	$p < 0.00001$	**0.239**
Off. vs PE+Sp	**59011**	**7.013**	$p < 0.00001$	**0.288**
Ggl vs GPT-b	73843	3.469	0.00043	0.142
Ggl vs GPT+Sp	**62409**	**6.201**	$p < 0.00001$	**0.254**
Ggl vs PE+Sp	**57082.5**	**7.474**	$p < 0.00001$	**0.307**
GPT-b vs GPT+Sp	75367	3.104	0.00162	0.127
GPT-b vs PE+Sp	68326.5	4.787	$p < 0.00001$	0.196
GPT+Sp vs PE+Sp	81396.5	1.664	0.0903	0.068

Table 4: Wilcoxon signed-rank test for translation pairs. A significant difference is defined as $p < 0.05$. Effect size (r) is interpreted as small (0.1), medium (0.3), and large (0.5).

These results show that ChatGPT PE + Spec is consistently preferred, in line with error-based evaluation findings. The low ranking of the official translation is notable, despite its presumed status as the gold standard. However, human translations often vary more than MT, depending on translator performance (Freitag et al., 2023; Ramos and Guzmán, 2024; Volz and von Thiessen, 2024). As a result, while the official translation was most frequently ranked in the lowest position, the number of times it was placed second, third, or fourth did not differ significantly. Moreover, it was ranked first more often than Google Translate and not far behind ChatGPT basic.

5.2.3 Qualitative Analysis

To gain insight into the stylistic and structural differences across translation types, we examine their sentence structure, focusing on relative clauses and clausal coordination. Our analysis shows that Google Translate and the official translations tend to use these forms more frequently, potentially reflecting source-language influence. Japanese allows for long, additive sentence constructions, which can lead to overuse of relative clauses or clausal coordination when translated too literally into English. Such structures may reduce readability, especially in English writing that values clarity and conciseness. Further details are provided in Appendix L.

We also analyze excerpts from the official translations and find recurring issues in grammar, style, and semantic clarity. For example, the expression "offering both a multitude of choices" contains a semantic mismatch between "both" and "multitude."

Type	Official	Google	GPT	GPT+Sp	GPT PE+Sp
Mean	*0.783*	**0.830**	0.822	0.821	0.810
SD	0.043	0.031	0.039	0.033	0.037

Table 5: Mean COMETKiwi scores and standard deviations for each translation type.

Other examples involved unidiomatic phrasing, sentence fragments, and inconsistent style. For detailed examples and qualitative error analysis, see Appendix M. These problems suggest that the low rating of the official translation may not stem from a lack of specifications but from variation in translator skill or mismatches with task requirements.

As mentioned in Section 5.2.2, human translations often vary in quality due to individual differences (Freitag et al., 2023; Ramos and Guzmán, 2024; Volz and von Thiessen, 2024). Combined with Japan's shortage of high-proficiency English translators (Appendix G), this may explain the observed results. By contrast, ChatGPT-based translations guided by specifications performed consistently well, suggesting their potential as a viable complement to traditional workflows.

5.3 Automatic Evaluation

We examine whether a reference-free automatic metric can capture differences in translation quality across specification and method types, compared to human judgment. To this end, we use COMETKiwi, a reference-free metric with the highest correlation to human evaluations in the WMT23 Metrics Shared Task (Rei et al., 2022; Freitag et al., 2023).[12] We adopt a reference-free approach because the official translations, typically used as references, are themselves part of the evaluation as one of the five translation types.

Table 5 shows scores from 0 to 1, with higher values indicating better quality. Low standard deviations suggest internal consistency, though overlapping ranges point to limited differences between types. Unlike the human rankings, COMETKiwi assigns the highest score to Google Translate and the lowest to the official translation. This divergence likely reflects differences in what COMET values, specifically literal fidelity and lexical similarity, as opposed to the more context-sensitive and stylistic qualities emphasized in our evaluation (Rei et al., 2022).

[12]We use the model `wmt22-cometkiwi-da`, also adopted as the WMT24 baseline for reference-free evaluation (Freitag et al., 2024).

Although the official translation appears to preserve source-like structures such as relative clauses and clausal coordination (Section 5.2.3, Appendix L), its low score may be partly explained by a few explanatory additions not present in the source, intended to assist international readers. Such additions may reduce source alignment and result in lower automatic scores.

ChatGPT PE + Spec scores slightly below Google Translate, though the difference is small. This may reflect Google Translate's more literal style, while ChatGPT PE + Spec balances fidelity and fluency, resulting in higher subjective appeal despite a lower COMET score. ChatGPT translations, particularly those guided by specifications, prioritize clarity and appeal over strict lexical matching, which COMET may not fully capture.

Although COMET metrics are known to struggle with numbers and named entities (Amrhein and Sennrich, 2022), our manual check found no significant errors in these areas, suggesting they did not affect the results.

As MT evaluation increasingly considers contextual and communicative goals, it is vital to develop automatic metrics that better capture functional aspects of translation quality, such as how well a translation fulfills its purpose in context.

6 Conclusion

We demonstrate that translation specifications can improve MT quality and enable more targeted evaluation. We provide a theoretical rationale for the importance of specifications, drawing on Skopos theory to support a functionalist perspective. Based on this foundation, we outline a practical guide for specification-aware MT using LLMs, including prompt design, generation, and both error-based and subjective evaluation.

In our case study, LLM outputs guided by specifications received higher ratings than official translations, Google Translate, or unguided LLM outputs. Although COMET scores favored Google Translate, they diverged from human evaluations. These findings suggest that specifications help LLMs produce more contextually appropriate translations that better align with communicative goals. The gap between human and automatic evaluations highlights the limitations of current metrics in capturing functional adequacy. Through this work, we demonstrate the potential of specification-aware MT for professional, real-world use cases.

Limitations

First, we use only a single LLM, ChatGPT. While its outputs are generally well-received, it occasionally introduces information not present in the source text. This highlights the importance of careful prompt design and human oversight, as is standard in professional translation workflows. Evaluating other LLMs remains an important area for future research to assess whether the findings generalize across models.

Second, our dataset consists of corporate philosophy statements from 33 Japanese companies, focusing solely on the Japanese-to-English language pair. While this allowed for a focused case study, broader validation will require larger datasets covering more diverse domains (e.g., legal and medical) and content types (e.g., marketing and technical manuals), as well as other language pairs.

Finally, our human evaluation process highlighted the difficulty of securing qualified annotators. Both the error analysis and the subjective evaluation were conducted through crowd-sourcing, and the compensation was set above the standard rates of that framework. For error analysis, which requires more specialized expertise, an alternative approach could have been to recruit evaluators through a more specialized platform and set the compensation accordingly. Such difficulties in recruiting and compensating qualified annotators emphasize the need to develop an automated and specification-based evaluation model. Future work could explore the *LLM as a Judge* (Zheng et al., 2023; Kocmi and Federmann, 2023b; Feng et al., 2025; Gunathilaka and de Silva, 2025), where an LLM evaluates outputs based on the same detailed specifications provided to human experts, potentially offering a scalable alternative to manual annotation.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful and constructive comments, which helped us improve this paper. This work was supported by JST FOREST Grant Number JPMJFR232R.

References

ISO 11669:2024. ISO 11669:2024. Translation projects – General guidance. https://www.iso.org/standard/79089.html. Accessed: 2025-07-29.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *The Third International Conference on Learning Representations*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M Guerreiro, Diptesh Kanojia, José GC de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, et al. 2023. Findings of the wmt 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Michael Carl, Akiko Aizawa, and Masaru Yamada. 2016. English-to-Japanese translation vs. dictation vs. post-editing: Comparing translation modes in a multilingual setting. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4024–4031, Portorož, Slovenia. European Language Resources Association (ELRA).

Sheila Castilho. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of*

the Workshop on Human Evaluation of NLP Systems (HumEval), pages 34–45, Online. Association for Computational Linguistics.

J. C. Catford. 1965. *Translation Shifts*, pages 73–82. Oxford University Press, London. Originally published as part of the series "Language and Language Learning" (Vol. 8).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Directorate-General for Translation (DGT), European Commission. 2015. DGT translation quality guidelines. `http://ec.europa.eu/translation/maltese/guidelines/documents/dgt_translation_quality_guidelines_en.pdf`.

Helge Dyvik. 1992. Linguistics and machine translation. In *Proceedings of the 8th Nordic Conference of Computational Linguistics (NODALIDA 1991)*, pages 67–78, Bergen, Norway. Norwegian Computing Centre for the Humanities, Norway.

ESG/Integrated Reporting Research Laboratory. 2024. 「統合報告書発行状況調査2023」最終報告（final report on the survey of integrated report publication status 2023）.

Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahan Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7084–7107, Vienna, Austria. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Paul Fields, Daryl Hague, Geoffrey Koby, Arle Lommel, and Alan Melby. 2014. What is quality? a management discipline and the translation industry get acquainted. *Tradumàtica: tecnologies de la traducció*, page 404.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Atsushi Fujita. 2021. Attainable text-to-text machine translation vs. translation: Issues beyond linguistic processing. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 215–230, Virtual. Association for Machine Translation in the Americas.

David Garvin. 1984. What does “product quality”’ really mean? *MIT Sloan Management Review*, 26:25–43.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Daniel Gouadec. 2007. *Translation as a Profession*, volume 73 of *Benjamins Translation Library*. John Benjamins Publishing.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

S. Gunathilaka and N. de Silva. 2025. Automatic analysis of app reviews using llms. In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 828–839. SciTePress.

Christian Hardmeier. 2015. On statistical machine translation and translation theory. In *Proceedings of the Second Workshop on Discourse in Machine Translation*, pages 168–172, Lisbon, Portugal. Association for Computational Linguistics.

Douglas Harper. 2025. explore | etymology, origin and meaning of explore by etymonline. Accessed: 2025-04-30.

Sui He. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 316–326, Sheffield, UK. European Association for Machine Translation (EAMT).

Bettina Hiebl and Dagmar Gromann. 2023. Quality in human and machine translation: An interdisciplinary survey. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 375–384, Tampere, Finland. European Association for Machine Translation.

Kurando Iida and Kenjiro Mimura. 2024. Cater: Leveraging llm to pioneer a multidimensional, reference-independent paradigm in translation quality evaluation.

ISO 18587:2017. 2017. Translation services – post-editing of machine translation output – requirements (iso standard no. 18587:2017).

ISO17100:2015. Translation services – requirements for translation services (iso standard no. 17100:2015).

ISO5060:2024. Translation services – evaluation of translation output – general guidance (iso standard no. 5060:2024).

Japan Exchange Group, Inc. 2021. Topix sector indices / topix-17 series.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Hui Jiao, Bei Peng, Lu Zong, Xiaojun Zhang, and Xinwei Li. 2024. Gradable chatgpt translation evaluation.

JTF. 2018. JTF翻訳品質評価ガイドライン第1版 (Japan Translation Federation translation quality evaluation guidelines first edition).

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2025. MQM-chat: Multi-dimensional quality metrics for chat translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.

Qibang Liu, Wenzhe Wang, and Jeffrey Willard. 2025. Effects of prompt length on domain-specific tasks for large language models.

Ting Liu, Chi-kiu Lo, Elizabeth Marshman, and Rebecca Knowles. 2024. Evaluation briefs: Drawing on translation studies for human evaluation of MT. In

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 190–208, Chicago, USA. Association for Machine Translation in the Americas.

Arle Lommel, Donald A. DePalma, and Tahar Bouhafs. 2024a. Q3 2024 language services market sizing update. Retrieved from https://csa-research.com.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024b. The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control. In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

A Melby. 2012a. Human and machine translation quality: Definable? achievable? desirable. In *LACUS Forum*, volume 39, pages 1–29.

Alan K. Melby. 2012b. Structured specifications and translation parameters (ttt.org specs). https://www.ttt.org/specs/#1b. Accessed: 2025-07-29.

Merriam-Webster Dictionary. 2025. Encounter. Accessed: 2025-04-30.

Ministry of Economy, Trade and Industry. 2022. Guidance for integrated corporate disclosure and company-investor dialogue for collaborative value creation 2.0.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

MQM. 2025. MQM (multidimensional quality metrics).

Jeremy Munday, Sara Ramos Pinto, and Jonathan Blakesley. 2022. *Introducing Translation Studies: Theories and Applications*, 5th edition. Routledge, London and New York.

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation.

Kazuaki Nagata. 2025. Japanese companies rush to up english-language disclosures in 2025. *The Japan Times*.

Eugene A. Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Brill, Leiden.

Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8568–8575.

Christiane Nord. 2006. Translating as a purposeful activity: A prospective approach. *TEFLIN Journal*, 17(2):131–143.

Haruka Ogawa. 2021. *Difficulty in English-Japanese translation: Cognitive effort and text/translator characteristics*. Ph.D. thesis, Kent State University.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson,

Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2021. Agree to disagree: Analysis of inter-annotator disagreements in human evaluation of machine translation output. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 234–243, Online. Association for Computational Linguistics.

Anthony Pym. 2023. *Exploring translation theories*. Routledge.

Fernando Prieto Ramos and Diego Guzmán. 2024. The impact of specialised translator training and professional experience on legal translation quality assurance: an empirical study of revision performance. *The Interpreter and Translator Trainer*, 18(2):313–337. PMID: 38812808.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Katharina Reiss and Hans J. Vermeer. 1984. *Grundlegung einer allgemeinen Translationstheorie*. Niemeyer, Tübingen. Translated into Spanish by S. García Reina and C. Martín de León as *Fundamentos para una teoría funcional de la traducción*, Madrid: Akal, 1996; translated into English by C. Nord and M. Dudenhöfer as *Towards a General Theory of Translational Action*, London and New York: Routledge, 2013.

Jie Ruan, Wenqing Wang, and Xiaojun Wan. 2024. Defining and detecting vulnerability in human evaluation guidelines: A preliminary study towards reliable NLG evaluation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7965–7989, Mexico City, Mexico. Association for Computational Linguistics.

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danielle Saunders. 2022. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *J. Artif. Int. Res.*, 75.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Mary Snell-Hornby. 1995. *Translation Studies: An Integrated Approach*, 2nd edition. John Benjamins, Amsterdam.

Mary Snell-Hornby. 2006. *The Turns of Translation Studies: New Paradigms or Shifting Viewpoints?* John Benjamins, Amsterdam.

Ingemar Strandvik. 2017. *Evaluation of outsourced translations. State of Play in the European Commission's Directorate-General for Translation (DGT)*. Language Science Press.

Runjia Tan, Xiang Long, and Oluwatoba O. Bamigbade. 2023. Comparative research on machine translation and human translation of examples in dictionary from the perspective of skopos theory. In *Proceedings of the 3rd International Conference on Internet, Education and Information Technology (IEIT 2023)*, pages 342–353. Atlantis Press.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Takeshi Ueno. 2025. Superior information communication and documentation. *The Worldfolio*.

Lawrence Venuti, editor. 2021. *The Translation Studies Reader*, 4th edition. Routledge, London and New York.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi,

United Arab Emirates (Hybrid). Association for Computational Linguistics.

J. P. Vinay and Jean Darbelnet. 1958. *Stylistique comparée du français et de l'anglais: méthode de traduction*, 1972 edition. Didier, Paris. Translated by J. C. Sager and M. J. Hamel as *Comparative Stylistics of French and English*, Amsterdam and Philadelphia, PA: John Benjamins, 1995.

Stephanie Volz and Raphael von Thiessen. 2024. Machine translation – recommendations for public administration. White paper. Canton of Zurich, Division of Business and Economic Development.

Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2023. Controlling styles in neural machine translation with activation prompt. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2606–2620, Toronto, Canada. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. Ai-assisted human evaluation of machine translation. In *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, New Mexico. Association for Computational Linguistics.

A Industry Trends from CSA Research

According to Lommel et al. (2024a) from CSA Research, an independent research firm specializing in the language services industry, language service providers (LSPs) that rely heavily on traditional human translation are experiencing declining performance. At the same time, while the growing demand for translation has outpaced the capacity of human translators, the increased use of MT alone has not led to sustainable growth (Lommel et al., 2024a).

The report encourages LSPs to shift their focus from merely producing high-quality translations to delivering greater value, such as providing cultural adaptation, adapting content for specific audiences, and training and customizing LLMs for domain-specific communication. It clearly states: "LSPs must focus on messaging that resonates with enterprise goals, and demonstrate that they use technology to achieve them (Lommel et al., 2024a)."

In response to these challenges, our study proposes a framework for specification-aware MT and evaluation.

B Historical Context of Controllable MT

As noted in Section 2.1, the goal of tailoring translation output is not new. In the eras of SMT and NMT, significant research focused on incorporating external knowledge to control specific aspects of translation.

For example, a major line of work involved leveraging existing human translations to improve consistency. This began with the convergence of Translation Memories and SMT (Koehn and Senellart, 2010) and was later adapted to NMT, such as through neural fuzzy repair mechanisms (Bulte and Tezcan, 2019). Other approaches focused on controlling discrete linguistic features, including methods to enforce terminology constraints during NMT decoding (Dinu et al., 2019) and to manage stylistic aspects such as formality (Niu and Carpuat, 2020).

While these approaches provided powerful control over discrete phenomena, they often required specialized data preparation or model retraining.

Our work builds upon this tradition but explores how prompt-based LLMs can manage a broader set of communicative specifications in a more flexible manner.

C Reliability of Evaluation Methods

Recent research has highlighted the need for more robust and transparent evaluation methods across NLP tasks, including but not limited to MT. This includes a growing interest in developing evaluation frameworks that are comprehensive, detailed, and interpretable. For instance, Nimah et al. (2023) propose the *Metric Preference Checklist*, an analytical framework that evaluates automatic natural language generation (NLG) metrics from five distinct perspectives, providing a more multifaceted evaluation of their alignment with human judgments.

In contrast, Xiao et al. (2023) point out that most research focuses only on how well metrics correlate with human ratings, often overlooking the reliability and measurement error of the metrics themselves. They argue that concepts from Measurement Theory, used in educational and psychological testing, should be applied to NLG evaluation to better assess the reliability and validity of evaluation metrics.

Gehrmann et al. (2023) outline a long-term agenda for improving NLG evaluation, including robust human evaluation protocols and the development of metrics that go beyond surface-level overlap. Meanwhile, Ruan et al. (2024) emphasize the low reliability of human evaluation guidelines in NLG, showing that only 29.84 percent of 3,233 papers in major NLP conferences shared their guidelines, and 77.09 percent of those contained some kind of vulnerability. They propose principles for more reliable guideline design and introduce a method using LLMs to detect guideline flaws.

These concerns are relevant to evaluation in MT. If the evaluation procedures are unclear, the reliability of the results cannot be ensured. In our study, we develop guidelines for both translation and evaluation and describe the experimental procedure based on these guidelines. In addition, instead of simply reporting the results, we offer analysis and possible interpretations for each finding to improve clarity and transparency.

D Garvin's Approach to Understanding Quality

Since the MQM framework is grounded in Garvin (1984)'s approach to quality, we provide a brief overview of his perspectives.

Garvin (1984), a prominent scholar in quality management, introduces five approaches to understanding quality: transcendent, product-based, user-based, manufacturing-based, and value-based. Among these, the manufacturing-based approach defines quality as meeting pre-set specifications, and the user-based approach emphasizes satisfying user needs.

Fields et al. (2014) discuss the importance of incorporating Garvin (1984)'s approach into translation quality assessment. Although the definition of translation quality is debated and Fields et al. (2014) disagree on some points, they generally agree that the production-based approach (originally "manufacturing-based" in Garvin's words), evaluating translation according to specifications, is important.

E Incorporating Specifications in MT and its Evaluation: The Case of MQM

Current research using MQM in MT evaluation often focuses on detailed error assessment. For example, the MQM framework has been used in shared tasks at the Conference on Machine Translation (WMT), such as the General Translation Task, the Metrics Task, and the Quality Estimation Shared Task.

However, specifications are not explicitly addressed in these tasks. This is because the purpose of the General Translation Task is to evaluate general MT capabilities, which may not require specific requirements, such as the purpose of the translation or the target audience, to translate the source text. This kind of general MT capability can be useful when users simply want to gain a general understanding of foreign content. However, if the goal is to communicate clearly in the target language, the translation must convey the message in a way that reflects its purpose and suits the intended reader. This cannot be done without specifications.

Freitag et al. (2021) report that in MQM evaluations of English-to-German and Chinese-to-English translations, approximately 80 percent of errors fall into the accuracy and fluency categories, with accuracy-related mistranslations being the most common. While accuracy and fluency are

relatively straightforward to assess, other error categories may be more difficult to judge without specifications. For example, the definition of the *Style* error is simply "Translation has stylistic problems" (Freitag et al., 2021). While this is distinct from *Fluency-grammar* errors, it may be difficult for evaluators to identify a stylistic issue if the purpose of the translation is not known.

In specification-based evaluation, if the purpose of the translation is to convey the cultural otherness of the source text and the specified style is a literal translation that closely follows the original, the translation should adhere to that style. In this case, the translation is expected to preserve the expressions and cultural markers of the source to maintain the visibility of the translation. In other words, even a natural and fluent translation may be considered an error if it minimizes the sense of translation under such specifications.

As mentioned earlier, specifications can also serve as a guide for evaluators, helping to reduce subjectivity in the evaluation process. In this sense, the detailed error types in MQM could be applied more effectively when used in conjunction with detailed specifications.

Several recent MQM-based approaches, such as GEMBA-MQM (Kocmi and Federmann, 2023a), Auto-MQM (Fernandes et al., 2023), xCOMET (Guerreiro et al., 2024), and MQM-APE (Lu et al., 2025), have laid important groundwork for the automation of fine-grained evaluation. However, translation specifications are not the central focus of these approaches.

While human evaluation approaches like MQM-Chat adapt MQM error categories to specific settings (e.g., chatbot applications), they do so by simply modifying the original MQM typology (Li et al., 2025). Their approach could be further extended to integrate detailed specifications. For example, it could require translations to preserve source-text ambiguity as a stylistic feature or to handle internet slang accurately as part of terminology, in line with MQM's original design.

These considerations highlight that MQM, while widely used, can be made more effective when applied in combination with translation specifications.

F A Practical Guide for Specification-Aware Machine Translation and Evaluation

We present guidelines for specification-aware translation and evaluation, drawing on the translation process in a typical professional workflow and the "Production Process" outlined in ISO 17100 —Requirements for Translation Services. This standard states that translation should be "in accordance with the purpose of the translation project, including the linguistic conventions of the target language and relevant project specifications" (ISO17100:2015).

Our evaluation method draws on ISO 5060, the JTF guidelines, and the MQM framework (ISO5060:2024; JTF, 2018; MQM, 2025).[13] Our approach combines these frameworks to make use of their different strengths in evaluation. MQM provides detailed error types for fine-grained analysis, while ISO 5060 and the JTF guidelines allow for weighted error categories based on specifications.[14] The JTF guidelines also emphasize the role of subjective evaluation (JTF, 2018). These evaluation guidelines are designed for use with both human and MT and assume that human evaluators will assess the output.

We first explain the method of specification-aware translation with prompt-based LLMs (Appendix F.1), followed by a description of the human error analysis procedure (Appendix F.2). It also addresses subjective evaluation, which provides a different perspective from error analysis for assessing whether the translation fulfills its intended purpose (Section F.3).

F.1 Specification-aware Machine Translations with Prompt-Based LLMs

Step 1: Define Translation Specifications Define detailed translation specifications. When working with clients or translation teams, all stakeholders should reach an agreement on the specifications. The following are examples corresponding to the parameters listed in Table 1, using a single translation project as context: the English translation of

[13]The European Commission's Directorate-General for Translation also provides quality evaluation guidelines based on a "fit for purpose" approach (Strandvik, 2017). Translations are assessed using error categories (e.g., Accuracy, Terminology, Linguistic Conventions, Style, and Formatting) with severity levels to calculate a quality score (Directorate-General for Translation (DGT), European Commission, 2015).

[14]The error classification used in ISO 5060 is based on the MQM framework (ISO5060:2024).

an integrated report published by a publicly listed Japanese company.

1. **Purpose of translation:** The translation's communicative goal (e.g., to inform, persuade, promote, or comply).

   Example: The goal is to attract foreign institutional investors. Therefore, the translation must emphasize growth potential and sustainability strategies in persuasive, investor-friendly language.

2. **Target audience:** The intended readers and their background knowledge, expectations, or needs.

   Example: The audience consists of non-Japanese institutional investors who may not be familiar with Japanese corporate structures. Key terms may require explanatory phrasing.

3. **Style, register and tone:** The desired level of formality, stylistic conventions, and voice appropriate to the target context.

   Example: A moderately formal and confident tone is preferred, neither overly technical nor overly casual, consistent with ESG reports by global competitors.

4. **Terminology and reference resources:** Required use of specific terms, glossaries, or past translations to ensure consistency.

   Example: The company previously translated 企業理念 as *Corporate Philosophy*, and this terminology should be maintained consistently across sections and future documents.

5. **Domain and legal requirements:** Industry-specific norms or legal constraints that affect wording or structure.

   Example: The report includes financial statements that must conform to IFRS terminology, and disclosures must reflect the Japan Financial Services Agency's guidelines on non-financial reporting.

6. **Cultural adaptation:** Modifications made to accommodate cultural expectations or sensitivities.

   Example: A phrase like *wa no seishin (*和の精神*)* may be unfamiliar to global readers and can be replaced with a culturally adapted equivalent such as *a spirit of harmony and mutual respect*.

7. **Length and formatting:** Constraints or allowances regarding text length, layout, or formatting elements.

   Example: Since the English translation must fit within the same layout as the Japanese version, sentence length and paragraph structure must be carefully managed to avoid overflow.

8. **Localization needs:** Adjustments for language variants, regional conventions, or localized preferences.

   Example: Dates, currencies, and units should follow international conventions (e.g., *FY2023* instead of *2023*年度, *million yen* instead of 百万円), and U.S. spelling is preferred for the target audience.

As these specifications demonstrate, creating appropriate translations requires numerous decisions. Defining these requirements in advance helps ensure that translations meet their intended purpose and are suitable for the target context.

Step 2: Define Roles Separate tasks performed by the machine, such as generating translations based on the given specifications, from those handled by humans, including supervision, quality assurance, verification of specification adherence, and the management of responsibilities such as meeting deadlines. Human reviewers oversee the overall process and focus on elements requiring expert judgment or domain-specific knowledge.

Step 3: Design Instructions Aligned with Specifications Create instructions for the LLM that reflect the defined specifications. Use the listed specifications to identify aspects that can be included in the model instructions. These instructions may include parameters such as:

- Source text information
- Target language
- Purpose of translation
- Target audience
- Style, register, and tone
- Length and formatting

These parameters may vary depending on the specific translation project. The instructions should also state that LLM must not add any information

that is not present in the original text. This is especially important in creative translation tasks, where the model may over-generate and introduce information not found in the source text.

Fine-tuning may be applied as needed, such as specifying terminology, aligning with existing translations, using consistent phrasing, maintaining preferred styles, or handling domain-specific vocabulary accurately.

Step 4: Generate Translation Use the instructions to generate the initial translation.

Step 5: Review and Finalization A human reviewer, or a team of reviewers, must carefully check whether the initial translation is accurate and ensure that it follows the defined specifications. Any errors should be corrected, and the translation should be finalized for delivery. This review process should be given sufficient time in the project schedule. It is standard practice to have reviewers check the translation to ensure quality and accuracy, even when the initial translation is done by professional human translators. The review helps identify issues that a single translator may miss. The same applies to LLMs.

F.2 Human Error Analysis Approach

Translation evaluation is not always conducted alongside the translation itself. There are various situations in which translation evaluation is required, such as deciding whether to accept or reject a translation, comparing multiple translation outputs, or selecting the most suitable version among candidates. Evaluation may also be needed for quality control in professional workflows, for assessing MT outputs, or for translator training and certification purposes.

Our error analysis methodology combines ISO 5060, the JTF guidelines, and the MQM framework to create a specification-aware evaluation system. This system is designed to score each translation based on how well it meets the predefined specifications. The more errors are found, the higher the score. Therefore, translations of higher quality should receive lower scores. By using these established frameworks, we develop a practical process that can be applied in various translation contexts.[15]

[15] Although full-text evaluation is ideal, practical constraints sometimes necessitate the use of samples. When translation samples are selected for evaluation, ISO 5060, MQM, and the JTF guidelines each offer instructions on how to carry out the sampling process (ISO5060:2024; MQM, 2025; JTF, 2018).

Step 1: Ensure Specifications Are Accessible Before evaluation begins, ensure that the translation specifications are available and accessible. If, for any reason, clear specifications are not defined at the time of translation, they should be established at this stage. The specifications should remain accessible throughout the evaluation process so that evaluators can refer to them consistently. Clear documentation helps ensure that all evaluators apply the same criteria and share a common understanding of the communicative goals of the translation.

Step 2: Define Error Categories Based on Specifications Establish error categories aligned with the specifications to ensure consistency. ISO 5060 defines a translation error as a "failure to adhere to translation project specifications" (ISO5060:2024).

However, it is important to distinguish between specifications and error categories. Specifications describe the requirements agreed upon for a project. In contrast, error categories cover a broader range of issues, including problems that are commonly assumed but not always stated directly in the specifications. For example, accuracy-related errors or violations of general linguistic conventions, such as grammar or punctuation, are typically included in error taxonomies even when they are not listed in the specifications. As long as the selected error categories do not contradict the specifications, they can be applied to support consistent evaluation.

Reference standards like the JTF guidelines and MQM for error categorization (JTF, 2018; MQM, 2025). MQM provides fine-grained error categories that are especially valuable for detailed error analysis (MQM, 2025).

Error levels can be customized based on project priorities. For example, when translating a company's corporate philosophy for investor relations materials, where conveying nuance and tone is more important than achieving word-for-word accuracy, the following adjustments can be made:

- Accuracy: Prioritize mistranslation errors, while placing less emphasis on over-translation and under-translation.
- Style: Use more specific subcategories, such as:
  - Language register: Inappropriate level of formality
  - Awkward phrasing: Grammatically correct but stylistically poor

  - Unidiomatic expressions: Unnatural to native speakers
  - Inconsistent style: Stylistic variations across the document

This targeted approach allows evaluators to focus on errors that are most relevant to the specifications. Error categories should be adjusted to reflect the specific requirements of each project.

Step 3: Apply Weights to Error Categories ISO 5060 and the JTF guidelines apply weights to error categories based on project specifications. These weights can be set in advance by the client or project owner. Since different document types prioritize different aspects of translation quality, stakeholders should agree on appropriate weights (ISO5060:2024; JTF, 2018):

- 2.0 – Highly Important
- 1.5 – Somewhat Important
- 1.0 – Standard Importance (default)
- 0.5 – Less Important

To maintain balance, the average weight across all categories should be approximately 1.0. The weighting example below places greater emphasis on accuracy and consistent currency formatting, while giving less weight to stylistic elements:

Example: Financial Report—Revenue Forecast Section

- Accuracy: 1.0 (Standard): Basic factual correctness is required, with some flexibility in expressing forecasts.
- Style: 0.5 (Less Important): A professional tone is preferred, but it has little impact on understanding.
- Locale Convention: 1.5 (More Important): All monetary values must be shown in US dollars to ensure a consistent interpretation.

Step 4: Select Qualified Evaluators Evaluators should be professional translators or subject matter experts who are not only bilingual but also native speakers of the target language. Being bilingual alone is not sufficient for proper evaluation; evaluators must have a deep understanding of the linguistic nuances and cultural context of the target language.

Regarding the qualifications and competencies of evaluators, the ISO 5060 and MQM frameworks provide more rigorous requirements (ISO5060:2024; MQM, 2025). Practitioners should refer directly to these standards to ensure that evaluators meet the necessary professional criteria.

However, these strict requirements can be challenging in practice, as it is often hard to find and recruit qualified evaluators. When qualified evaluators are difficult to recruit, it may be useful to combine error-based evaluation conducted by available bilingual reviewers with subjective assessment by domain experts in the target language or end users. Domain experts may detect inconsistencies or errors by closely reading the content, while end users can provide direct feedback on whether the translation feels natural or conveys the intended message.

Step 5: Identify and Assess Errors Qualified evaluators identify errors, record them, and assess their severity. Severity indicates how serious an error is. It should always be judged based on whether the translation achieves its intended purpose and how much real-world impact the error may have (JTF, 2018). If an error has little or no practical impact, assigning a severity level may not be necessary.

Severity levels and their corresponding scores follow JTF (2018):

- Neutral (0): No penalty. These include stylistic preferences or repeated minor issues that do not affect comprehension.
- Minor (1): Errors that slightly affect readability but do not interfere with understanding.
- Major (10): Errors that significantly affect readability and comprehension.
- Critical (100): Errors that make the translation unusable and may cause harm, such as health risks, financial losses, or reputational damage. These must be corrected before publication.

Whether to count repeated errors multiple times should be decided through agreement among stakeholders (JTF, 2018).

Step 6: Calculate the Score Each identified error receives a score calculated as:

$$\text{Error Score} = \text{Category Weight} \times \text{Severity Score} \quad (1)$$

After assessing all errors, sum the scores to calculate the total error score. If severity scoring is not used, simply multiply the number of errors in each category by its assigned weight, and then sum the results.

A lower total score indicates fewer errors, and therefore a higher-quality translation. These scores make it possible to directly compare different translations.

For projects that require pass/fail decisions, evaluators can set a passing threshold based on acceptable error levels. Refer to the JTF guidelines or the MQM framework for recommended threshold values that fit different translation contexts (JTF, 2018; MQM, 2025).

As the evaluation is based on detailed specifications, it allows for a more objective assessment, reducing subjectivity and ensuring consistent criteria among evaluators.

F.3 Human Subjective Evaluation Approach

The JTF guidelines emphasize that while error-based evaluation methods provide a systematic approach, they represent only one aspect of translation quality (JTF, 2018). Error-based methods focus primarily on objectively identifiable errors and do not account for subjective quality factors that are essential in certain types of documents, such as advertisements, literary works, corporate vision statements, marketing slogans, and brand messages. For a more comprehensive assessment of quality, it is advisable to combine error-based evaluation with other approaches, particularly subjective evaluation by experts or end users, depending on the translation context (JTF, 2018).

Moreover, when qualified evaluators for error-based assessment are unavailable, subjective evaluation can complement error analysis performed by available bilingual reviewers. Subjective evaluation captures aspects such as clarity, persuasiveness, and appropriateness, which are essential for determining whether a translation effectively serves its intended purpose.

Subjective evaluation can be integrated with error-based assessment in the following ways:

A. Subjective Evaluation by Experts : Experts assess translations based on their specialized knowledge and professional judgment. Examples of such expert evaluations include:

- Legal professionals assess the accuracy and appropriateness of legal translations.
- Marketing specialists review the effectiveness and cultural relevance of promotional content.
- Technical staff on the client side evaluate clarity, precision, and technical correctness in the documentation.

B. Subjective Evaluation by End Users : End users evaluate translations based on their own perceptions and practical experience. For example, in the case of investor relations materials, investors may be asked:

- *Did you find the explanation clear, convincing, and appropriate?*
- *Were there any unnatural expressions or confusing elements in the translation?*

Their feedback is usually collected through surveys or questionnaires and provides valuable insight into how clear and usable the translation is.

Incorporating specifications into translation and evaluation enables both to go beyond basic accuracy and focus on communicative effectiveness. This ensures translations are not only correct but also appropriate for their intended audiences and contexts.

G Why Focus on Japanese-to-English Translation

Our experiment focuses on Japanese-to-English translation of integrated reports, which are typically published annually by companies. Here, we explain why this particular language pair and translation direction were chosen.

MT research often prioritizes universal approaches, aiming to develop models and evaluation methods that generalize across many language pairs. For example, Liu et al. (2024) note that this focus on generalization is evident in the field's pursuit of standardized methods. While such approaches help improve general performance across languages, they may overlook challenges specific to individual language pairs.

Indeed, prior research has reported substantial differences in MT performance between high-resource and low-resource languages (Team et al., 2022; Pang et al., 2025). These disparities have been attributed not only to the quantity of training data but also to inherent linguistic factors. For instance, Sarti et al. (2022) found that post-editing greatly improved English–Italian translations, but

had limited impact on English–Turkish and English–Japanese, likely due to word order and morphology differences. Similarly, Lee et al. (2022) report that mBART (Tang et al., 2021) performs well across domains but struggles with typologically distant languages, scoring below 3.0 BLEU (Papineni et al., 2002).

In response, our study takes a more targeted perspective, recognizing that translation difficulty varies widely depending on linguistic distance, grammatical structure, and cultural context. Japanese–English translation presents unique challenges due to fundamental differences in linguistic structure and writing systems. Unlike English–European language pairs, which often require minimal restructuring, Japanese–English translation typically involves major changes in sentence structure and word choice (Carl et al., 2016). Ogawa (2021) states that translating between Japanese and English takes significantly more processing time than English–German or English–French translation, leading to a higher cognitive load.

Due to these complexities, general approaches to translation models and evaluation, particularly those designed for multilingual settings, may not fully reflect the specific challenges of Japanese–English translation. Therefore, a specialized approach is necessary to understand not only how MT systems handle these linguistic difficulties, but also how their outputs can be appropriately evaluated.

We focus on the Japanese-to-English translation direction for two main reasons. First, ChatGPT and other LLM-based translation systems tend to perform better in English than in many other languages, due to the abundance of high-quality English training data (Chowdhery et al., 2022; Wendler et al., 2024). For example, OpenAI reports that GPT-3's training data is "primarily English (93 [percent] by word count)," and similar English-centric characteristics are reflected in the design and evaluation of GPT-4, as noted in its System Card (Brown et al., 2020; OpenAI et al., 2024). Working in this direction allows for both practical translation and evaluation while minimizing the influence of data-related limitations.

Second, the demand for Japanese-to-English translation greatly exceeds the supply of qualified human translators. Japan continues to face a shortage of professionals who can produce high-quality English translations, especially in specialized fields such as investor relations and corporate communications (Nagata, 2025; Ueno, 2025). Given this shortage, MT guided by translation specifications represents a practical alternative that may help meet the demand for high-quality and cost-effective translations. In this study, we examine whether this approach can improve translation quality while also addressing the translator shortage and meeting international communication needs.

H Company List

Table 6 lists the integrated reports used in this study. One company is selected from each of the 33 industry categories defined by the Tokyo Stock Exchange, prioritizing those ranked among the top four in market capitalization within each category during the data collection period (August 28–September 9, 2024).

The table includes the industry, company name, and publication year of the report used. Major companies with larger market capitalizations are chosen under the assumption that they are more likely to publish well-developed English versions of their integrated reports. As a result, many of the companies listed are well-known Japanese corporations.

I Prompt Design for Specification-Aware ChatGPT Translations

Table 7 summarizes the prompts used for each ChatGPT translation method. For ChatGPT basic, ChatGPT receives only a minimal instruction: "Please translate the following Japanese text into English." For ChatGPT + Spec, we incorporate the content of the assumed specifications for the translation of integrated reports:

- **Source text context**: The official company name and a description of the integrated report as an IR document
- **Target language**: English
- **Intended purpose**: To enhance the company's appeal to a broad audience of investors
- **Target audience**: International investors
- **Style**: Clear and persuasive, suitable for a global investor audience

In the ChatGPT PE + Spec method, the model is instructed to improve the Google Translate out-

Industry	Company	Year
Transportation Equipment	Toyota Motor Corp.	’23
Banks	Mitsubishi UFJ Financial Group, Inc.	’23
Electric Appliances	Sony Group Corp.	’23
Retail Trade	Fast Retailing Co., Ltd.	’23
Services	Recruit Holdings Co., Ltd.	’23
Information & Communication	Nippon Telegraph and Telephone Corp.	’23
Chemicals	Shin-Etsu Chemical Co., Ltd.	’24
Wholesale Trade	Mitsubishi Corp.	’23
Pharmaceuticals	Chugai Pharmaceutical Co., Ltd.	’23
Other Products	ASICS Corp.	’23
Insurance	Tokio Marine Holdings, Inc.	’23
Foods	Japan Tobacco Inc.	’23
Precision Instruments	Terumo Corp.	’23
Fishery, Agriculture & Forestry	Nippon Suisan Kaisha, Ltd.	’23
Mining	Japan Petroleum Exploration Co., Ltd.	’23
Construction	Daiwa House Industry Co., Ltd.	’23
Textiles & Apparels	Goldwin Inc.	’23
Pulp & Paper	Oji Holdings Corp.	’24
Oil & Coal Products	ENEOS Holdings, Inc.	’23
Rubber Products	Bridgestone Corp.	’24
Glass & Ceramics Products	AGC Inc.	’24
Iron & Steel	Nippon Steel Corp.	’23
Nonferrous Metals	Sumitomo Electric Industries, Ltd.	’23
Metal Products	Sanwa Holdings Corp.	’23
Machinery	Mitsubishi Heavy Industries, Ltd.	’23
Electric Power & Gas	The Kansai Electric Power Co., Inc.	’23
Land Transportation	Central Japan Railway Co.	’23
Marine Transportation	Nippon Yusen Kabushiki Kaisha	’23
Air Transportation	All Nippon Airways Co., Ltd.	’23
Warehousing & Harbor Transportation Services	Mitsui-Soko Holdings Co., Ltd.	’23
Securities & Commodity Futures	Nomura Holdings, Inc.	’24
Other Financing Business	Japan Exchange Group, Inc.	’23
Real Estate	Mitsui Fudosan Co., Ltd.	’23

Table 6: Integrated reports of 33 listed companies used in the experiment.

put based on the same specifications, without access to the original Japanese text. This is a deliberate design choice; providing the source text risks the model disregarding the MT output and producing a new translation from scratch, a phenomenon observed in our initial pilot experiments. We therefore ensured that the task remained genuine post-editing, focused solely on enhancing the fluency and appeal of the existing translation. In both specification-aware methods, we instruct ChatGPT to avoid adding information not present in the source.

Since this is not a commissioned project, we define realistic specification parameters based on industry practices. Items like deadlines and formatting are excluded due to experimental constraints.

J Example Comparison of Five Translation Methods: All Nippon Airways Co., Ltd.

The five translations in Table 8 illustrate different approaches to conveying the original Japanese text in English. The source text is an excerpt from the corporate philosophy section of the integrated report published by All Nippon Airways Co., Ltd.

The official translation retains the Japanese term “waku waku,” along with an explanatory note. It is unclear whether the original specifications required preserving the Japanese phrase. However, even if the intention was to reflect a sense of Japanese cultural identity, the primary objective should be to ensure that the translation appeals to international investors. One notable issue is that in the phrase “[a]nd when passed from person to person, becomes a force,” the subject is missing, which could be considered an error. Additionally, the phrase “explore more incredible discoveries” may sound awkward, as “explore” typically refers to something unknown rather than something already discovered (Harper, 2025).

Google Translate output is simple and easy to understand, but expressions such as “make travel more fun” and “more exciting sky experiences” might be too casual and do not fit the context of

Translation Type	Prompt
ChatGPT basic	Please translate the following Japanese text into English.
ChatGPT + Spec	The following Japanese text is an excerpt from the integrated report of [company name], a key part of the company's investor relations materials. Please translate this text into English in a way that will be appealing to international investors. The purpose of this translation is to enhance the company's appeal to a wider audience of investors. Please do not add any additional information.
ChatGPT PE + Spec	The following text is a translation of an excerpt from the integrated report of [company name], a key part of the company's investor relations materials. The purpose of this translation is to enhance the company's appeal to a wider audience of investors. The initial translation was done using Google Translate. Please refine this translation to make it more engaging and appealing in English. Please do not add any additional information.

Table 7: Examples of ChatGPT prompts used in our study.

Type	Translation Output
Source	"ワクワク"は, 人を動かすエネルギー. それは人から人へと伝わり、世界をあかるく元気にする。私たちは, 空の可能性を信じ, ヒト・モノ・コトをつなぐことで, これまでたくさんの笑顔と歓びに出会ってきた. そして今, 新しい未来に向けて, 次のワクワクを生み出す. 移動をもっと楽しいものにできないか, モノの運び方を新しくできないか, 心躍る日常を届けられないか, どこまでも地球にやさしくなれないか. これまで以上の空の体験とともに, 驚きと感動を創りたい.
Official	"Waku waku" is what moves people to push what's possible. It's Japanese for the joy and excitement of discovering the unknown. And when passed from person to person, becomes a force that creates a brighter world, united in wonder. The sky is full of possibilities, which has allowed us to bring together people, products and experiences for decades. Now, we're expanding our horizons and spreading "waku waku" across the world. So, we wonder, can we make travel more enjoyable? Can we renew the way we transport products? Can we bring excitement to life, while being kind to the earth? Together with an elevated experience in the sky, we can explore more incredible discoveries that will lead to a better tomorrow.
Google	"Excitement" is the energy that moves people. It spreads from person to person, making the world brighter and more energetic. We believe in the possibilities of the sky, and by connecting people, things, and events, we have encountered many smiles and joy. And now, we are creating the next excitement for a new future. Can we make travel more fun, can we innovate the way we transport things, can we deliver an exciting daily life, and can we be kind to the earth as much as possible? We want to create surprises and excitement along with more exciting sky experiences than ever before.
GPT	"Excitement" is the energy that moves people. It spreads from person to person, bringing brightness and vitality to the world. Believing in the possibilities of the sky, we have connected people, goods, and experiences, encountering countless smiles and joy along the way. Now, we are setting out to create the next source of excitement for a new future. Can we make travel more enjoyable? Can we reinvent the way goods are transported? Can we deliver everyday moments that make your heart dance? Can we be more environmentally friendly than ever before? Along with new and enhanced experiences in the sky, we aim to create wonder and inspiration.
GPT+Sp	Excitement is the energy that moves people. It spreads from person to person, brightening and invigorating the world. We believe in the limitless potential of the skies and, through connecting people, goods, and experiences, we have encountered countless smiles and moments of joy. Now, as we look toward a new future, we are ready to create the next wave of excitement. Can we make travel more enjoyable? Can we innovate the way things are transported? Can we bring a sense of wonder to everyday life? Can we go further in protecting our planet? We aim to create surprises and inspiration with an even greater sky experience.
GPT PE+Sp	"Excitement" is the spark that moves people, spreading from one person to another, brightening and energizing the world. Believing in the boundless potential of the skies, we bring people, goods, and experiences together, creating countless moments of joy and countless smiles. Now, we are embarking on a new journey to inspire even greater excitement for the future. Can we make travel more delightful, revolutionize the way we transport goods, infuse everyday life with excitement, and care for our planet in the best possible ways? We are committed to creating moments of surprise and delight, offering more thrilling experiences in the skies than ever before.

Table 8: Differences in translations: All Nippon Airways Co., Ltd.

investor relations materials. The phrase "encountered many smiles and joy" also sounds unnatural, as "encounter" is typically used with concrete entities or situations, such as difficulties or opposition, rather than with abstract concepts like joy or smiles (Merriam-Webster Dictionary, 2025).

ChatGPT basic captures the emotional aspects of the original text, particularly with phrases like "moments that make your heart dance." However, the use of "encountering countless smiles and joy," similar to Goolge Translate, does not sound natural in English.

ChatGPT + Spec employs expressions such as "limitless potential of the skies," which are effective in creating a positive and aspirational tone. However, the use of "things" in "the way things are transported" sounds casual, and "even greater sky experience" would sound more natural if "experiences" were used in the plural form.

ChatGPT PE + Spec employs strong and active expressions. The word "spark" in "[e]xcitement is the spark that moves people" creates a vivid and powerful impression, while the phrase "embarking on a new journey to inspire even greater excitement" conveys a positive and future-oriented feeling.

These observations suggest that each translation method tends to produce distinct results, and that including specifications in ChatGPT prompts may encourage the use of more purposeful and engaging language.

K Error Typology for Human Error Evaluations

We use the following categories and subtypes for error annotation: Accuracy (subtypes: mistranslation, addition, omission), Linguistic Conventions (grammar, spelling, unintelligible, textual conventions), and Style (language register, awkward style, unidiomatic style, inconsistent style). Definitions and examples of each error type are available on the official MQM website (`https://themqm.org/downloads/`). Evaluators are instructed to refer to this table during error analysis. However, only the main categories are used for error counting; subtypes are provided to help annotators better understand and identify specific issues.

L Syntactic Pattern Analysis of Translation Outputs

To investigate stylistic tendencies, we analyzed syntactic patterns across the five translation types, focusing on relative clauses and clausal coordination. Specifically, we counted the number of relative pronouns (*which*, *who*, *that*) and clausal instances of *and* (i.e., those connecting two clauses with subject-verb structures). We excluded uses of *that* as complementizers, demonstratives, or in cleft constructions, and excluded *and* used at the phrase or word level.

These structures are common in Japanese texts and may reflect a literal transfer of source syntax. Their frequent use can lead to complex, additive structures that may reduce readability in English.

Table 9 reports total word counts, raw counts, and normalized frequencies per 1,000 words. The results show that Google Translate and the official translations use relatively more relative clauses and clause-level coordination, suggesting less restructuring. In contrast, ChatGPT outputs display simpler sentence structures regardless of prompt specificity. These patterns indicate that prompt-based LLMs tend to favor fluency and conciseness.

M Examples and Error Analysis from Official Translations

To explore why the official translations received low ratings, we examined problematic excerpts from companies' corporate philosophies. Table 10 presents these examples.

To begin with, Excerpt (1) contains the phrase "offering both a multitude of choices," in which the use of *both* appears semantically inappropriate. The word *both* typically introduces two parallel elements, but "a multitude of choices" is a singular, collective concept, resulting in a semantic mismatch. In this context, *both* is presumably intended to refer to the two entities mentioned earlier, *individuals* and *businesses*. However, its placement creates a structurally awkward and confusing expression.

Excerpt (2) opens with the sentence "[w]hat we do isn't a job," which appears to aim for an inspirational tone but lacks a clear referent or elaboration. As a result, its meaning may be ambiguous to readers who are not familiar with the intended message behind the expression. A clearer alternative might be "[w]hat we do is more than a job," which conveys the intended message more directly.

Excerpt (3) contains the phrase "in this era of search," which is not a commonly used expression in English discourse and may not be immediately clear. Additionally, the combination of the degree

Type	Official	Google	ChatGPT	ChatGPT+Sp	ChatGPT PE+Sp
Word Count	8,776	**8,894**	8,655	8,351	8,216
Clausal *and*s	58	**66**	54	53	57
Rel. Pronouns	72	**83**	61	59	61
*and*s / 1000w	6.61	**7.42**	6.24	6.35	6.94
RelP / 1000w	8.20	**9.34**	7.05	7.07	7.42

Table 9: Syntactic feature counts per translation type. RelP = relative pronouns (*which*, *who*, and *that*). Frequencies normalized per 1,000 words.

No.	Excerpt
(1)	Opportunities for Life. Faster, simpler and closer to you. Since our foundation, we have connected individuals and businesses, offering both a multitude of choices. (Recruit Holdings Co., Ltd.)
(2)	What we do isn't a job. We enjoy exploring what is possible for our future. We question the status quo, fail well and overcome with resilience. We are a force for change. (Recruit Holdings Co., Ltd.)
(3)	In this era of search, where information has become available anytime anywhere, we need to focus more on proposing the optimal choice. We seek to provide 'Opportunities for Life,' much faster, surprisingly simpler and closer than ever before. (Recruit Holdings Co., Ltd.)
(4)	Today, what we mean by Our Hopes for the Future, a world where we are our truest selves, respecting, and inspiring each other. Living together in harmony with our planet—in harmony with People and Nature. (Daiwa House Industry Co., Ltd.)
(5)	Create a virtuous cycle between Society and Earth by fully utilizing less of her limited resources. Make the world a richer, better place by bringing out the best out in people and the potential of buildings. (Daiwa House Industry Co., Ltd.)
(6)	For the sake of the Earth, which future generations of children have entrusted in our care. Together with you. (Bridgestone Corp.)
(7)	The single continuous curve represents the dynamism and our commitment for continuous innovation and delivering value to people and society. (Nippon Telegraph and Telephone Corporation)

Table 10: Excerpts from the official translation cited in the qualitative analysis.

adverb *surprisingly* with the comparative adjective *simpler* creates a stylistic inconsistency.

Excerpt (4) is grammatically incomplete. The subject and predicate do not form a complete clause, making the intended meaning difficult to determine.

Excerpt (5) uses the phrase "fully utilizing less," which appears to aim for a concise message about efficiency, likely meaning "to make the most of fewer resources." However, the expression is semantically ambiguous. The adverb *fully* suggests maximization, while *less* implies minimization, creating a tension that may confuse readers rather than clarify the company's commitment to sustainability. In addition, the phrase "bringing out the best out in people" is grammatically incorrect. The structure redundantly includes both "out" before and after "the best," where only one instance is appropriate. A corrected version would be "bringing out the best in people," which is idiomatic and clear.

Excerpt (6) contains grammatical issues. The phrase "entrusted in our care" is unidiomatic. In standard English, the verb *entrust* typically appears in the form "entrust someone with something" or entrust something to someone." In addition, the sentence lacks a main clause and does not constitute a complete grammatical unit.

Finally, Excerpt (7) contains two issues. First, "commitment for" is a grammatical error. The standard preposition in this context is "commitment to." Second, the coordination of "continuous innovation," a noun phrase, and "delivering value," a gerund phrase, is unbalanced and stylistically awkward. For clarity and parallel structure, both elements should be in the same grammatical form, such as "continuous innovation and value creation," or "innovating continuously and delivering value."

As these examples show, the official translation includes not only grammatical inaccuracies but also semantic and stylistic inconsistencies, which may have contributed to its lower rating in the evaluation.

The characteristics of the source texts themselves may help explain the relatively low ratings of the official translations. Integrated reports sometimes contain expressions that are abstract, culture-specific, or metaphorical in Japanese, which can result in awkward or even ungrammatical output if translated too literally. In some cases, the official translations appear to reflect such overly direct translations, suggesting that the translator may have prioritized fidelity to the source text's wording or syntax at the expense of naturalness and clarity in English. While this approach may have been inten-

tional, for example to preserve a uniquely Japanese tone, it can hinder readability and reduce the overall appeal of the translation. This is reflected in the lower rankings observed in the subjective evaluation.

However, it is important to note that professional translations are typically produced based on specifications or an internal guideline (ISO17100:2015). In practice, translators are expected to follow such instructions from the outset; without them, it would be difficult to even begin the task. Therefore, the types of problems identified in the official translations, such as grammatical errors or awkward phrasing, are unlikely to stem from missing or unclear specifications. Rather, these issues may be related to language proficiency or a mismatch between the translator's background and the specific requirements of the task.

Specifications can support translation decisions, but achieving linguistic accuracy and fluency may still require a high level of language proficiency. As noted earlier (Section 5.2.2), human translation quality tends to vary, as translators differ in background and ability (Freitag et al., 2023; Ramos and Guzmán, 2024; Volz and von Thiessen, 2024).

Furthermore, as discussed in Appendix G, Japan continues to face a shortage of translators capable of producing high-quality English translations. This shortage may have influenced the present results.

In this context, the potential of MT systems that use specifications, such as ChatGPT with customized prompts or post-edited outputs, deserves more attention. Both ChatGPT + Spec and ChatGPT PE + Spec are favorably evaluated in our study, not only in subjective rankings but also in error-based analysis, suggesting that specification-aware MT may offer a useful complement to traditional workflows.

OpenWHO: A Document-Level Parallel Corpus for Health Translation in Low-Resource Languages

Raphaël Merx[λ] **Hanna Suominen**[ψ,ϕ] **Trevor Cohn**[λ] **Ekaterina Vylomova**[λ]
[λ] School of Computing and Information Systems, The University of Melbourne
[ψ] School of Computing, The Australian National University
[ϕ] School of Medicine and Psychology, The Australian National University

Abstract

In machine translation (MT), health is a high-stakes domain characterised by widespread deployment and domain-specific vocabulary. However, there is a lack of MT evaluation datasets for low-resource languages in this domain. To address this gap, we introduce OpenWHO, a document-level parallel corpus of 2,978 documents and 26,824 sentences from the World Health Organization's e-learning platform. Sourced from expert-authored, professionally translated materials shielded from web-crawling, OpenWHO spans a diverse range of over 20 languages, of which nine are low-resource. Leveraging this new resource, we evaluate modern large language models (LLMs) against traditional MT models. Our findings reveal that LLMs consistently outperform traditional MT models, with Gemini 2.5 Flash achieving a +4.79 ChrF point improvement over NLLB-54B on our low-resource test set. Further, we investigate how LLM context utilisation affects accuracy, finding that the benefits of document-level translation are most pronounced in specialised domains like health. We release the OpenWHO corpus to encourage further research into low-resource MT in the health domain.

1 Introduction

Translation in the health domain combines clinical risks, widespread demand, and domain-specific complexity (Mehandru et al., 2022; Neves et al., 2024). By offering a timely and resource-efficient complement to human translation, machine translation (MT) can lower the barrier to disseminating health content, from education materials for local health workers (Hammond et al., 2024) to public safety information during crises (Federici et al., 2023; Utunen et al., 2023b). However, evaluation of MT in the health domain is hampered by a lack of datasets that cover a wide range of languages, particularly low-resource ones. The TICO-19 corpus (Anastasopoulos et al., 2020) stands as a notable exception, yet its focus on COVID-19 limits its utility on broader health topics, and its age raises the risk of training data contamination.

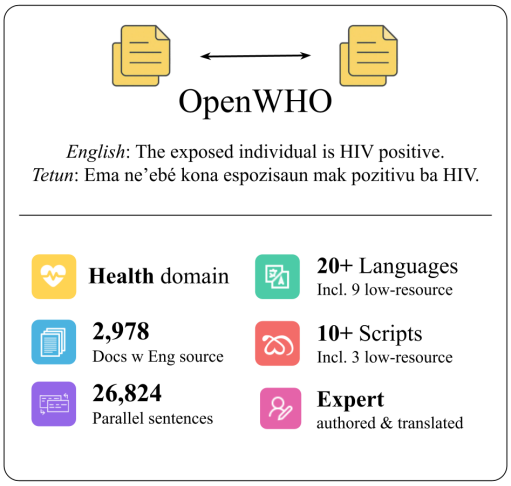


Figure 1: Overview of the OpenWHO parallel dataset, highlighting its depth across low-resource languages and scripts.

To address this gap, we introduce OpenWHO, a document-level parallel corpus designed for evaluating health MT. Sourced from the World Health Organization's multilingual e-learning platform, its content is expert-authored, professionally translated, and shielded from web-crawling, thus minimising contamination risk. The corpus covers over 20 languages, nine of which are low-resource, including some with low-resource scripts like Armenian, Georgian, and Sinhala. By focusing on health education, a domain fundamental to local quality of care (Merx et al., 2024b), OpenWHO provides a realistic benchmark for a high-impact MT use case.

Leveraging this new resource, we conduct a systematic evaluation comparing modern large language models (LLMs) against traditional NMT systems. For LLMs, we study different context strategies (document-level, sentence-level, etc) and to

Proceedings of the Tenth Conference on Machine Translation, pages 142–160
November 8-9, 2025

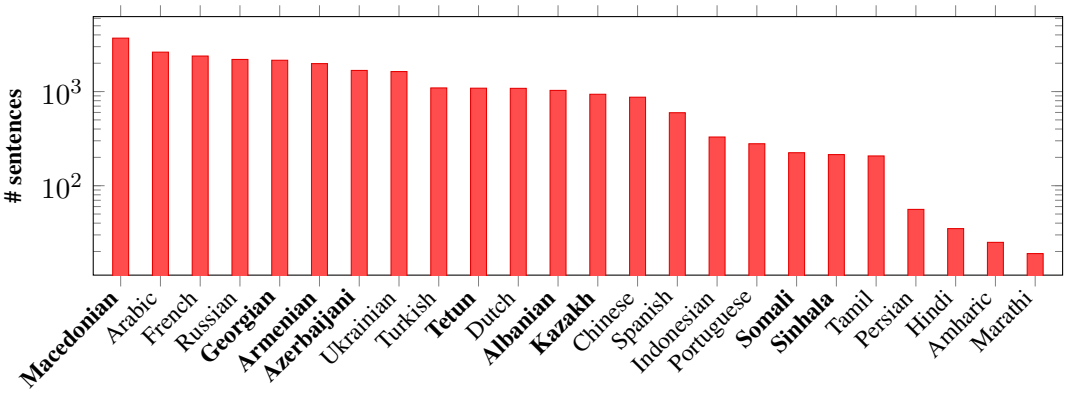


Figure 2: Number of parallel sentences per language in the OpenWHO dataset. The English source has 50,898 sentences. Low-resource languages covered in our experiments (Section 4) are in **bold**.

determine whether the benefits of document-level translation are specific to our dataset, we extend our evaluation to the news and literary subsets of the WMT24++ benchmark (Deutsch et al., 2025).

Our main contributions are [1]:

- **We introduce and release OpenWHO**, a parallel corpus for health MT that covers low-resource languages. It comprises 2,978 documents and 26,824 parallel sentences from expert-authored, professionally translated materials (§3).
- **We show that modern LLMs outperform traditional NMT models** for low-resource translation in the health domain. Our findings show Gemini 2.5 Flash with document-level context achieves a +4.79 ChrF point improvement over NLLB-54B on our test set (§4.2).
- **We find that the benefit of document-level context is model and domain-dependent** for low-resource MT. Accuracy gains are most pronounced when using best-performing models to translate specialised domains like health and literature, while the general (news) domain shows more modest improvements, highlighting that domain complexity drives context utility (§4.2).

2 Related Work

Document-level MT Document-level MT has long been recognised as desirable, as it allows models to leverage broader discourse for improved coherence and accuracy (Maruf et al., 2021). Early work with traditional NMT models has shown mixed results, with some studies demonstrating that document-level context can significantly improve translation quality (Miculicich et al., 2018; Wang et al., 2023; Post and Junczys-Dowmunt, 2024), while others questioned whether these improvements stemmed from true contextual understanding, arguing that the context encoder was not modeling discourse but acting as a "noise generator" that improves model robustness, rather than leveraging discourse information (Li et al., 2020; Appicharla et al., 2024).

LLMs for document-level MT LLMs, given their ability to process extended contexts, are well placed to benefit from document-level context. Koneru et al. (2024) explored contextual translation with Llama2-13B on English-German, finding mixed results, where document-level context sometimes providing no performance gains over sentence-level translation. Karpinska and Iyyer (2023) demonstrated that paragraph-level translation outperforms sentence-level approaches in literary fiction using GPT-3.5, though they have noted that their findings might not generalise to low-resource settings. Yang et al. (2024) fine-tuned LLaMA3-8B for context-aware translation, showing that increasing context window size yields gains, particularly when evaluated with neural metrics. Recent mechanistic analysis by Mohammed and Niculae (2025) revealed that LLMs can be surprisingly "context-insensitive," with smaller models showing limited ability to effectively utilise available context, a finding that may explain varying performance observed in earlier works.

Low-resource MT with LLMs While LLMs have shown promising results on low-resource MT

[1] Code: github.com/raphaelmerx/openwho-code
Dataset: huggingface.co/datasets/raphaelmerx/openwho

Work	Low-resource	LLMs	Document-level	Specialised domains
Ours	✓	✓	✓	✓
Post and Junczys-Dowmunt (2024)	✗	✗	✓	✗
Wang et al. (2023)	✗	✗	✓	✗
Koneru et al. (2024)	✗	✓	✓	✗
Karpinska and Iyyer (2023)	✗	✓	✓	✓
Enis and Hopkins (2024)	✓	✓	✓	✗
Zebaze et al. (2025)	✓	✓	✗	✓
Yang et al. (2024)	✗	✓	✗	✓
Mohammed and Niculae (2025)	✗	✓	✓	✗
Pang et al. (2025)	✗	✓	✓	✓

Table 1: Comparison with prior work on context utilisation for MT.

(Guo et al., 2024; Merx et al., 2024a), evaluation has been predominantly limited to sentence-level translation. Enis and Hopkins (2024) demonstrated that Claude significantly outperforms NLLB on Yoruba-English translation, finding substantial improvements from document-level over sentence-level translation, though their evaluation focused solely on the low-resource-to-English direction. Zebaze et al. (2025) explored low-resource translation with LLMs using compositional approaches on datasets including FLORES, NTREX, and TICO-19, but operated at the sentence level with few-shot learning rather than document-level context. This sentence-level focus in low-resource settings represents a significant gap, as the dynamics of context utilisation may be fundamentally different for low-resource languages where models have seen limited training data and where translation errors could compound across sentences within a document.

Gaps remain in our understanding of document-level translation for low-resource languages in specialised domains. First, there is a shortage of evaluation data for document-level low-resource machine translation in specialised domains, such as healthcare. Second, there has been no systematic analysis of how LLMs utilise document-level context when translating low-resource languages, particularly in specialised domains where coherence and terminological consistency are required.

3 The OpenWHO dataset

3.1 Source and Motivation

The OpenWHO platform. Our corpus is drawn from OpenWHO.org, the World Health Organization's (WHO) former e-learning platform for public health education. Active from 2017 to 2024, the platform's primary goal was to disseminate health knowledge to healthcare professionals, frontline responders, and the public, particularly during health emergencies (George et al., 2022; Utunen et al., 2023a). The content was authored and vetted by WHO experts and its global network of partner institutions, ensuring that the information and its translations were authoritative, accurate, and reflected up-to-date scientific guidance (George et al., 2022). The topics covered a wide range of public health issues, including specific disease responses (e.g., COVID-19, Ebola), vaccination protocols, infection prevention, and emergency preparedness (Utunen et al., 2020, 2023a).

Multilingual focus. A key tenet of the OpenWHO initiative was to ensure equitable access to information, which included a deliberate strategy of multilingual dissemination. Course materials were translated from English into a range of languages, with a focus on providing resources for low- and middle-income countries (George et al., 2022; Utunen et al., 2023a). The course-based format ceased operations in December 2024, transitioning to a static resource library. The data for our corpus was collected prior to this change and exclusively comprises materials from the course-based period (2017–2024). This commitment to creating expert-authored, multilingual content made the OpenWHO platform a high-quality source for extracting a document-level parallel corpus in the health domain, covering several low-resource languages.

Because all course material was hosted behind a login screen, it was shielded from the large-scale web crawling that constitutes the training data for most LLMs, mitigating risk of pre-training contamination. To confirm this, we conducted searches across publicly available web-scraped corpora (C4, MADLAD), and performed targeted web searches (via Google Search) using sentences found in OpenWHO course content. These searches revealed only publicly accessible OpenWHO course descriptions

(which are not part of our corpus), with no course content found within these data sources.

3.2 Data Curation Pipeline

3.2.1 Document Extraction

Scraping While we secured authorization from the WHO to collect and release this data, a direct database export was not available. Therefore, in consultation with the WHO, we developed a web scraping pipeline to gather the course materials. Using the Scrapy framework,[2] we developed a web scraper to navigate the OpenWHO site, enrol in each individual course, and extract the raw HTML content of every course page. Each page was uniquely identified by its course ID and language, as well as its position within the course structure (section and subsection numbers).

Content filtering. A significant portion of the OpenWHO curriculum relies on video-based learning. As our focus is on creating a parallel text corpus, we filtered out pages where video was the primary medium. To further ensure the quality of the extracted documents, we applied a series of heuristic filters to remove low-value content: we discarded pages that primarily consisted of a list of references, contained fewer than ten words, or featured boilerplate text used to introduce a course or section.

3.2.2 Document Pairing

The structured nature of the OpenWHO platform facilitated document alignment. For any given course page, the quadruplet (`language code`, `course id`, `section index`, `subsection index`) serves as a unique identifier. By varying the `language code`, we could accurately identify and group parallel course pages that are direct translations of one another.

After applying the quality filters described in the previous section, this pairing process yielded $2,978$ parallel documents. This set includes significant coverage for several low- and mid-resource languages, with Tetun, Albanian, Macedonian, Azerbaijani, Kazakh, Georgian, and Armenian each having more than 50 parallel documents. A breakdown of document counts per language is presented in Table 5.

[2] https://www.scrapy.org/

3.2.3 Sentence Mining

Traditional NMT models rely on sentence-level translation. To release a dataset that can be used for NMT evaluation, and potentially fine-tuning, we mined parallel sentences from parallel documents.

Annotation To evaluate our sentence mining pipeline, we manually annotated 10 parallel documents for 8 target low-resource languages (Macedonian, Georgian, Armenian, Azerbaijani, Tetun, Albanian, Kazakh, Tamil) that are of interest to our experiment. For each document pair, we manually segmented the source and target texts into aligned sentences (relying on back-translation for the languages we are not familiar with), ensuring one-to-one correspondence. This process yielded a reference corpus totalling $2,645$ parallel sentences. This annotated set serves as the ground truth for the subsequent sentence-splitting and alignment evaluation.

Sentence splitting Using the target-language sentences from our manually annotated corpus as a reference, we evaluated the performance of three sentence-splitters: `NLTK` (Bird and Loper, 2004), `Stanza` (Qi et al., 2020), and `pysbd` (Sadvilkar and Neumann, 2020). We measured sentence splitting performance as accuracy of sentence boundaries against our ground truth segmentation. The results (shown in Appendix B) indicated that `pysbd` achieves the highest accuracy overall with accuracy ranging from 82.0% for Kazakh to 94.0% for Tetun, but `stanza` performs better for Kazakh (89.1%), Tetun (94.6%) and Georgian (93.2%). NLTK's performance was generally lower than the other two. Based on these findings, we selected the best-performing tool (either `pysbd` or `stanza`) on a per-language basis to segment the entire corpus.

Sentence alignment After sentence splitting, we aligned sentences to create parallel pairs. Here we rely on sentence semantic similarity, using LaBSE (*Language-agnostic BERT Sentence Embedding* (Feng et al., 2022)), which supports all our languages of interest except Tetun (as a consequence, for Tetun, we first translated target sentences back to English before encoding). Because the OpenWHO documents are relatively short, this approach is highly effective: when evaluated against our manually annotated ground truth, the method yielded F1 scores ranging from 98.6% (for Tetun) to 100% (for Kazakh and Georgian).

Strategy	Description
Sentence level	Our baseline. Each sentence is translated independently without any additional context.
Sentence window (batched sliding window in Koneru et al. (2024))	A constrained-context approach. The model receives only the immediately preceding and succeeding sentences as context, aiming to capture local discourse phenomena without overwhelming the model.
Sentence + doc context	The model is provided with the full source document as context within the prompt but is instructed to translate only the single, target sentence.
Document level	The model is given the entire source document and instructed to translate the whole text. As per Enis and Hopkins (2024), we use one sentence per line, and evaluate at the sentence level after translation.
Doc-level + self-correct, as per (Wu et al., 2025)	A two-step approach: (1) Document-level translation then (2) feed the generated translation back to the model with a new prompt asking it to review and improve its own output, testing its self-correction and refinement capabilities.

Table 2: The five translation strategies evaluated in our experiments. Each strategy represents a different approach to leveraging context for machine translation. Associated prompts are in Appendix F.

Quality Control and Filtering Finally, to ensure the quality of the mined sentence pairs, we implemented an additional filtering stage based on empirical rules. We removed sentence pairs that were likely to be misaligned or uninformative for translation tasks. This included removing (1) pairs where the source English sentence contained fewer than five words, as these are often section headers or fragments; (2) pairs where the target-language side was in English; and (3) sentence pairs that were exact duplicates across different course pages, which often correspond to repeated instructions or boilerplate phrases.

Starting from an initial pool of $43,732$ candidate sentence pairs, we arrived at a final, clean set of **26,824** parallel sentences. This includes nine low-resource languages with over 200 parallel sentences (Macedonian, Georgian, Armenian, Azerbaijani, Tetun, Albanian, Kazakh, Somali, Sinhala). The count of sentence pairs per language is detailed in Table 5.

3.3 Dataset Statistics

The resulting OpenWHO corpus comprises **2,978** parallel documents and **26,824** aligned parallel sentences between English and over 20 other languages. The corpus contains a mix of high-resource and low-resource languages, with significant depth in the latter, including six with over 1,000 parallel sentences (Macedonian, Georgian, Armenian, Albanian, Kazakh, and Tetun). A key feature of this dataset is its origin: all content is expert-authored and professionally translated, providing high-fidelity, domain-specific text that is a level above standard web-crawled corpora in terms of quality and consistency. The data is structured at both the document and sentence level, enabling experiments in document-level machine translation, terminology extraction, and domain adaptation. However, a potential weakness of this dataset is its unbalanced language distribution, as not all courses were translated into all languages.

3.4 Data Availability

With permission from the WHO, we release this dataset under a Creative Commons NonCommercial license (CC BY-NC 4.0), allowing re-use, modification and distribution for non-commercial use, while requiring attribution. Data will be available both at the document level and at the sentence level.

4 Experiments

Having established an evaluation corpus for document-level low-resource MT in the health domain, we now turn to investigating what models perform best on this dataset, and how context utilisation strategies affect LLM performance on this dataset. Our experimental design addresses a fundamental tension in document-level translation: while broader context can improve coherence and terminological consistency, it may also introduce noise or lead to error propagation.

We work with the following research questions:

- **RQ1**: How do state-of-the-art LLMs compare to traditional NMT models for health low-resource translation?
- **RQ2**: What is the most effective context strategy for LLM-based translation into

low-resource languages? (sentence-level, document-level, sliding sentence window, etc)

- **RQ3**: How does model capability interact with these context strategies?

4.1 Experimental Setup

Datasets We evaluate on two datasets, always in the EN-XX direction. The first is our newly introduced OpenWHO corpus. To ensure a controlled comparison across languages, we focus our experiments on a single, extensively translated course: "Infection Prevention and Control through Hand Hygiene (IPC-HH)". We select the nine low- to mid-resource languages available for this course for our evaluation: Albanian (sqi), Armenian (hye), Azerbaijani (aze), Georgian (kat), Kazakh (kaz), Macedonian (mkd), Sinhala (sin), Somali (som), and Tetun (tet). For a comparison with high-resource languages, we separately evaluate on French (fra), Russian (rus) and Spanish (spa).

To test the generalisability of our findings beyond the health domain, we also evaluate on the WMT24++ benchmark (Deutsch et al., 2025), an expansion of the WMT24 dataset to 55 languages. To align with our research focus, we select a sample of five low- to mid-resource languages present in this dataset: Bulgarian (bul), Serbian (srp), Swahili (swh), Tamil (tam), and Zulu (zul). Because this dataset is available at the paragraph level, for our sentence-level analysis, we split paragraphs into aligned sentences using Gemini 2.5 Flash.

Models Our model selection is designed to compare modern LLMs (both open and closed weights) against conventional NMT baselines. For NMT baselines, we select NLLB-200 (3.3B & 54B, Costa-jussà et al., 2024) and MADLAD-400 10B (Kudugunta et al., 2023), both of which cover languages covered in our evaluation (except Tetun for NLLB). For LLMs, we select Gemini 2.5 Flash (Gemini Team, 2025), a powerful closed-weight model, DeepSeek-V3 671B (DeepSeek-AI, 2024), which represents the state-of-the-art in open-weight models at the time of our experiments, and Gemma 3 27B (Team, 2025), a smaller LLM with broad multilingual support. We run all model calls through OpenRouter.[3]

Metrics We primarily evaluate with ChrF++ (Popović, 2017), an n-gram based metric which has been shown to correlate better with human judgement than BLEU (Papineni et al., 2002) particularly for morphologically rich languages like Kazakh or Georgian. To validate results found with ChrF++, we also evaluate with MetricX-24[4] (Juraska et al., 2024) and AutoMQM (Fernandes et al., 2023). MetricX is a neural metric which correlates better with human judgement than ChrF++ for high-resource languages. While it has not been evaluated on low-resource languages, it is based on mT5 (Xue et al., 2021), which has been pretrained on all languages in our study, aside from Tetun. AutoMQM uses a large language model to characterise translation errors using MQM (Lommel et al., 2013). To avoid self-preference bias that may arise from using the same LLM for AutoMQM as that used for translation (Wataoka et al., 2025), we run AutoMQM with Kimi K2 (Kimi-AI, 2025).

	ChrF ↑	MetricX ↓
OpenWHO (9 low-res langs)		
NLLB 54B	50.52	3.45
Gemini	55.32 ↑ **4.79**	3.10 ↓ **-0.43**
DeepSeek-v3	49.38 ↓ **-1.14**	3.92 ↑ **0.39**
Gemma 3	48.01 ↓ **-2.51**	4.24 ↑ **0.71**
WMT24++ literary (5 low-res langs)		
NLLB 54B	43.00	5.83
Gemini	50.66 ↑ **7.66**	3.76 ↓ **-2.07**
DeepSeek-v3	46.88 ↑ **3.88**	4.57 ↓ **-1.26**
Gemma 3	44.45 ↑ **1.45**	5.26 ↓ **-0.57**
WMT24++ news (5 low-res langs)		
NLLB 54B	53.58	3.45
Gemini	54.83 ↑ **1.24**	2.69 ↓ **-0.76**
DeepSeek-v3	51.40 ↓ **-2.18**	3.42 ↓ **-0.04**
Gemma 3	50.71 ↓ **-2.87**	3.61 ↑ **0.16**

Table 3: Average performance per model, with score difference from NLLB 54B. Modern LLMs like Gemini outperform NLLB on specialised domain low-resource MT, like health or literary fiction. See Tables 7 and 10 for scores per language, which vary from 37 to 63 ChrF.

Translation Strategies For LLM translation, we rely on a fixed one-shot prompt (Appendix F), and we systematically evaluate five translation strategies that introduce contextual information in different ways. Detailed in Table 2, these include translating sentences one at a time, translating sentences with some surrounding context, and translating entire documents at once. For NMT models (NLLB

[3] https://openrouter.ai/

[4] google/metricx-24-hybrid-large-v2p6-bfloat16

and MADLAD), as they were trained at the sentence level, we evaluate only at the sentence level. To ensure a fair comparison across models and strategies, all outputs, including those generated at the document level, are segmented and evaluated at the sentence level against the reference translations.

4.2 Results

LLMs outperform NMT on health low-resource translation (RQ1). On OpenWHO, Gemini 2.5 Flash, when translating at the document level, outperforms NLLB 54B across all languages,[5] by an average of +4.79 ChrF points (Table 3). MetricX and AutoMQM results confirm this overall trend. However, other LLMs evaluated (DeepSeek-v3 and Gemma 3) are still outperformed by NLLB-54B, albeit by a small margin for DeepSeek. This means that among open weight models, NLLB-54B is still the preferred choice. Further, at equivalent performance before fine-tuning, LLMs require far more computation, with around one order of magnitude more parameters for the same performance (e.g. DeepSeek-v3 671B roughly equivalent to NLLB 54B; Gemma 3 27B equivalent to NLLB 3.3B).

Error analysis: Gemini vs NLLB Error analysis using AutoMQM (Table 12) shows that Gemini translations contain substantially fewer critical errors than NLLB, with less mistranslations (where target text does not accurately represent the source meaning) and less incorrect terminology, at the cost however of more omissions (where target text is missing information present in the source) and overtranslations (target text more specific than the source).

On high-resource languages, NLLB and LLMs are very close to each other. On our sample of high-resource OpenWHO languages (French, Russian, Spanish), the average scores for NLLB 54B, Gemini, DeepSeek, and Gemma 3 are all remarkably close to each other, as measured by both ChrF (averages in the 59-62 range for all 4 models) and MetricX (averages in the 2.3-2.4 range). This result indicates that in the health domain, the advantage of LLMs over NMT is more pronounced on low-resource languages compared to high-resource. Unsurprisingly, performance on high-resource languages is notably higher than on low-resource ones, with a gap of 7-12 ChrF points between high-resource and low-resource across all models.

LLMs tend to work best at the document level, for specialised domains (RQ2). On OpenWHO, both Gemini and DeepSeek translate best at the document level, with +3.62 and +2.00 ChrF points over sentence-level translation respectively (Table 4), but no measurable improvements in MetricX scores. On WMT24++ literary, the advantage of document-level over sentence-level is even clearer, with +6.37 Chrf points for Gemini, +3.34 for DeepSeek, and similar improvements in MetricX scores (Table 11). For Gemma 3 27B however, additional context from document-level only marginally improves translation accuracy, on both OpenWHO and WMT24++ literary. Overall, we observe a trend where **the larger the LLM, the more it benefits from document-level translation (RQ3)** over sentence-level translation.

In the general domain, the advantage of modern LLMs and document-level translation are less clear. On the WMT24++ news set, we do not see meaningful accuracy improvements for document-level over sentence-level translation, using either metric (ChrF and MetricX). We also see less variation in scores between models on this domain, both when comparing NLLB to LLMs, and when comparing LLMs with each other. Overall, the advantage of document-level translation over sentence-level translation for low-resource MT is not uniform across domains and models.

5 Discussion

Our experiments present **three key findings**: First, modern LLMs tend to outperform NMT (e.g. Gemini outperforms NLLB 54B) on low-resource translation in specialised domains (health with OpenWHO, literary text with WMT24++). Second, modern LLMs translate best at the document-level in specialised domains (health and literary), but the advantage of document-level translation is less clear for smaller models and for the general domain. Third, other context-utilisation strategies (e.g. sentence window, document context with one sentence at a time) tend to perform less well than whole-document translation.

Why Gemini outperforms NLLB in low-resource specialised domain MT (RQ1) Our investigation into performance differences between

[5] We exclude Tetun from this comparison, as NLLB does not support it.

Doc vs sent	ChrF Δ	MetricX Δ
OpenWHO (9 low-res langs)		
Gemini	↑ 3.62	→ 0.00
DeepSeek	↑ 2.00	→ 0.02
Gemma3	↓ -0.21	↑ 0.24
WMT literary (5 low-res langs)		
Gemini	↑ 6.37	↓ -1.18
DeepSeek	↑ 3.34	↓ -0.79
Gemma3	↑ 2.06	↓ -0.14
WMT news (5 low-res langs)		
Gemini	↑ 1.24	↓ -0.08
DeepSeek	↓ -0.82	↓ -0.11
Gemma3	↓ -0.14	↑ 0.13

Table 4: Performance difference for document-level vs sentence-level translation, averaged across languages. In specialised domains (health, literary fiction), the larger the LLM, the more it benefits from doc-level translation. See Tables 8 and 11 for scores per language.

LLMs and NMT models reveals that Gemini's advantage over NLLB 54B in specialised domains stems directly from its ability to leverage document-level context. When both models are constrained to sentence-level translation, their performance is very similar across all three datasets evaluated (OpenWHO, WMT24++ literary, and WMT24++ news, all within a narrow 1.5 ChrF point margin). It is only when Gemini is provided with the full document that it establishes a clear performance lead.

The role of context strategy across domains Our findings show that the optimal context strategy depends on both text domain (RQ2) and model capability (RQ3). The benefit of document-level translation is most pronounced in specialised domains like health and literature, potentially because their discourse structure requires a high degree of linguistic coherence for both accuracy (e.g. correct health terminology) and stylistic integrity (e.g. sustained narrative tone). In contrast, the news domain may rely more on self-contained sentences that allow skimming and quoting, reducing the benefit of context. Further, our results indicate that smaller models only gain marginal benefits from document context, potentially lacking the capacity to maintain coherence without introducing noise.

Our findings, particularly the dependence of context utility on model capability and domain specificity, offer a nuanced picture for where document-level context is most useful, which may explain past work that either did not (Li et al., 2020; Appicharla et al., 2024; Koneru et al., 2024) or did (Wang et al., 2023; Post and Junczys-Dowmunt, 2024; Wu et al., 2024) find added benefits from contextual level translation.

Recommendations Based on our results, we offer three recommendations for researchers working on low-resource MT:

1. **Evaluate LLMs at the document level for specialised domains.** Sentence-level evaluation can mask the advantage of modern LLMs, which lies in their ability to use context.

2. **Utilise the most capable LLMs to maximise the benefit of document context.** The performance gains from document-level translation are most significant with the largest models.

3. **Analyse performance on a per-language basis.** Average model rankings do not always reflect performance on individual languages, making granular analysis essential for model selection.

Future directions Several avenues for future work emerge from our findings. First, the development of reliable evaluation metrics tailored to low-resource MT in the health domain. Second, further exploration of strategies to optimise LLM-based translation for low-resource health contexts, such as fine-tuning on domain-specific data or different prompting techniques. Third, the creation of evaluation benchmarks for low-resource health on other tasks, such as question answering, which OpenWHO could be leveraged for.

6 Conclusion

In this work, we introduced OpenWHO, a high-quality parallel corpus for health MT, with a focus on low-resource languages. Sourced from the World Health Organization's expert-authored materials, it addresses a gap in evaluation resources and provides a benchmark for future research at the intersection of health and low-resource languages. The dataset strengths include (1) the grounding of its source English text in evidence-based WHO guidance (2) its professional translation into various languages and (3) its availability at both the document and sentence level. However, OpenWHO is language imbalanced (not all courses were translated into all languages), which can limit its comparative value.

Our experiments demonstrate that modern LLMs, when provided with full document-level context, outperform traditional NMT models on low-resource translation in specialised domains like health and literature. We found that this advantage is most pronounced for the largest models and diminishes in the general (news) domain, highlighting that the utility of context depends on both model capability and domain complexity. Our work underscores the potential of document-aware LLMs to improve translation quality in high-impact settings, while also revealing the critical need for domain-specific evaluation benchmarks and context-aware translation strategies.

7 Limitations

Metrics Our findings rely exclusively on automated metrics (ChrF, MetricX, AutoMQM). While these metrics give a useful signal when they all agree, we have limited ability to resolve differences when they arise. ChrF is a recognised standard for low-resource MT but may not always correlate well with human judgement (Wang et al., 2024); MetricX and AutoMQM have not been evaluated on low-resource languages, let alone in the health domain. Overall, more work is needed to determine what is the right metric for low-resource health MT, including a comprehensive human evaluation to validate our findings and gain a more nuanced understanding of translation quality.

Generalisability across other domains In our experiments on context utilisation, we rely on two specialised domains: health (OpenWHO) and literary fiction (WMT24++). While we find similar trends, our findings may not generalise to other specialised domains, such as legal, financial, or technical texts. The specific characteristics of each domain may influence the utility of document-level context, and a broader, structured evaluation across multiple domains would be needed to draw more general conclusions.

Caveats of a direct comparison between LLMs and NMT While document-level LLM translation beats sentence-level NMT translation for the languages and specialised domains we evaluate on, this comparison might be unfair to NMT models, which could be adapted to benefit from document-level context for a more equivalent comparison, and have far fewer model parameters at equivalent performance levels. In practice, LLM outputs could be leveraged for knowledge distillation, creating smaller, domain-specific models that retain much of the performance advantage while being more efficient (Gibert et al., 2025).

Dataset language imbalance Finally, the OpenWHO dataset itself has limitations. Its language distribution is imbalanced, as not all source materials were translated into every target language. This can constrain its utility for direct cross-language comparisons.

8 Ethics Statement

Consent This work adheres to ethical guidelines for data collection and research in natural language processing. The OpenWHO corpus was compiled from the WHO's e-learning platform with explicit authorization from the WHO for both data collection and public release. Our work aligns with the WHO's mission to disseminate health information globally and respects their ownership of the content.

Dual use and societal impact We have carefully considered the potential for dual use of the OpenWHO dataset and our research findings. Our primary objective is to enhance access to health education material by improving MT for low-resource languages in the high-stakes health domain. The dataset comprises expert-authored, professionally translated public health materials, limiting risks of misuse. The humanitarian and public health benefits of facilitating information access in underserved languages significantly outweigh dual-use concerns.

Acknowledgements

We are deeply grateful to the World Health Organization (WHO) for their collaboration and for granting us permission to collect and publicly release the OpenWHO dataset. In particular, we would like to express our sincere gratitude to Heini Utunen, Corentin Piroux, and Melissa Attias for their support and guidance on this project.

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann,

Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Ramakrishna Appicharla, Baban Gain, Santanu Pal, Asif Ekbal, and Pushpak Bhattacharyya. 2024. A case study on context-aware neural machine translation with multi-task learning. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 246–257, Sheffield, UK. European Association for Machine Translation (EAMT).

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846. Publisher: Nature Publishing Group.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *arXiv preprint*. ArXiv:2502.12404 [cs].

Maxim Enis and Mark Hopkins. 2024. From LLM to NMT: Advancing Low-Resource Machine Translation with Claude. *arXiv preprint*. ArXiv:2404.13813 [cs].

Federico M. Federici, Christophe Declercq, Jorge Díaz Cintas, and Rocío Baños Piñero. 2023. Ethics, Automated Processes, Machine Translation, and Crises. In Helena Moniz and Carla Parra Escartín, editors, *Towards Responsible Machine Translation: Ethical and Legal Considerations in Machine Translation*, pages 135–156. Springer International Publishing, Cham.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Richelle George, Heini Utunen, Ngouille Ndiaye, Anna Tokar, Lama Mattar, Corentin Piroux, and Gaya Gamhewage. 2022. Ensuring equity in access to online courses: Perspectives from the WHO health emergency learning response. *World Medical & Health Policy*, 14(2):413–427. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wmh3.492.

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling Low-Resource MT via Synthetic Data Generation with LLMs. *arXiv preprint*. ArXiv:2505.14423 [cs].

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Robert Hammond, Antonito Hornay Cabral, Jeremy Beckett, Xhian Meng Quah, Natarajan Rajaraman, Sanjay Mathew, Amrutha Gopalakrishnan, Mariano Pereira, Manuel Natercio Noronha, Bernardo Pinto, João de Jesus Arcanjo, Celia Gusmao dos Santos, Telma Joana Corte-Real de Oliveira, Ingrid Bucens, and Charlotte Hall. 2024. Lessons Learnt Delivering a Novel Infectious Diseases National Training Programme to Timor-Leste’s Primary Care Workforce. *Annals of Global Health*, 90(1).

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Kimi-AI. 2025. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual Refinement of Translations: Large Language Models for Sentence and Document-Level Post-Editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and Safe Use of Machine Translation in Medical Settings. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2016–2025, New York, NY, USA. Association for Computing Machinery.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024a. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Raphaël Merx, Christine Phillips, and Hanna Suominen. 2024b. Machine Translation Technology in Health: A Scoping Review. In *Health. Innovation. Community: It Starts With Us*, pages 78–83. IOS Press.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2025. Context-Aware or Context-Insensitive? Assessing LLMs' Performance in Document-Level Translation. *arXiv preprint*. ArXiv:2410.14391 [cs].

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the Classic: Revisiting Challenges of Machine Translation in the Age of Large Language Models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2024. Escaping the sentence-level paradigm in machine translation. *arXiv preprint*. ArXiv:2304.12959 [cs].

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Heini Utunen, Ranil Appuhamy, Melissa Attias, Ngouille Ndiaye, Richelle George, Elham Arabi, and Anna Tokar. 2023a. Observations from three years of online pandemic learning response on OpenWHO.

The International Journal of Information and Learning Technology, 40(5):527–540. Publisher: Emerald Publishing Limited.

Heini Utunen, Ngouille Ndiaye, Corentin Piroux, Richelle George, Melissa Attias, and Gaya Gamhewage. 2020. Global Reach of an Online COVID-19 Course in Multiple Languages on OpenWHO in the First Quarter of 2020: Analysis of Platform Use Data. *Journal of Medical Internet Research*, 22(4):e19076. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Heini Utunen, Thomas Staubitz, Richelle George, Yu Ursula Zhao, Sebastian Serth, and Anna Tokar. 2023b. Scale Up Multilingualism in Health Emergency Learning: Developing an Automated Transcription and Translation Tool. *Studies in Health Technology and Informatics*, 302:408–412.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2025. Self-Preference Bias in LLM-as-a-Judge. *arXiv preprint*. ArXiv:2410.21819 [cs].

Di Wu, Seth Aycock, and Christof Monz. 2025. Please Translate Again: Two Simple Experiments on Whether Human-Like Reasoning Helps Translation. *arXiv preprint*. ArXiv:2506.04521 [cs].

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting Large Language Models for Document-Level Machine Translation.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xinye Yang, Yida Mu, Kalina Bontcheva, and Xingyi Song. 2024. Optimising LLM-Driven Machine Translation with Context-Aware Sliding Windows. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1004–1010, Miami, Florida, USA. Association for Computational Linguistics.

Armel Zebaze, Benoît Sagot, and Rachel Bawden. 2025. Compositional Translation: A Novel LLM-based Approach for Low-resource Machine Translation. *arXiv preprint*. ArXiv:2503.04554 [cs].

A Documents and sentences per language

Language	Script	Number of documents	Number of sentences
Russian (rus)	Cyrillic	315	2194
French (fra)	Latin	301	2385
Arabic (ara)	Arabic	293	2623
Macedonian (mkd)	Cyrillic	254	3695
Ukrainian (ukr)	Cyrillic	204	1632
Chinese (zho)	Chinese	203	871
Spanish (spa)	Latin	149	596
Georgian (kat)	Georgian	131	2151
Armenian (hye)	Armenian	125	1982
Kazakh (kaz)	Cyrillic	103	936
Azerbaijani (aze)	Latin	98	1677
Turkish (tur)	Latin	81	1093
Indonesian (ind)	Latin	80	329
Dutch (nld)	Latin	77	1082
Albanian (sqi)	Latin	74	1029
Tetun (tet)	Latin	67	1086
Portuguese (por)	Latin	62	279
Hindi (hin)	Devanagari	28	35
Tamil (tam)	Tamil	26	207
Sinhala (sin)	Sinhala	25	214
Persian (fas)	Perso-Arabic	25	56
Amharic (amh)	Ethiopic	22	25
Marathi (mar)	Devanagari	21	19
Somali (som)	Latin	20	224
Italian (ita)	Latin	20	60
Lao (lao)	Lao	18	29
Yoruba (yor)	Latin	17	14
Burmese (mya)	Burmese	15	32
Swahili (swa)	Latin	14	107
Vietnamese (vie)	Latin	11	13
Catalan (cat)	Latin	10	8
Pushto (pus)	Perso-Arabic	8	22
Hausa (hau)	Latin	8	28
Thai (tha)	Thai	7	8
Shan (shn)	Shan	7	3
S'gaw Karen (ksw)	Karen	7	2
Japanese (jpn)	Japanese	6	9
Bulgarian (bul)	Cyrillic	6	9
Bengali (ben)	Bengali	6	8
Urdu (urd)	Perso-Arabic	4	10
Telugu (tel)	Telugu	4	4
Greek (ell)	Greek	4	4
Serbian (srp)	Latin	3	4
Polish (pol)	Latin	3	5
Panjabi (pan)	Gurmukhi	3	4
Oriya (ori)	Odia	3	4
Kurdish (kur)	Latin/Arabic	3	3
Tajik (tgk)	Cyrillic	2	1
Romanian (ron)	Latin	2	6
Nigerian Pidgin (pcm)	Latin	2	–
Lingala (lin)	Latin	1	7

Table 5: Number of OpenWHO documents and sentences per language. Low-resource languages are in **bold**.

B Sentence splitting performance

Method	Tamil	Armenian	Azerbaijani	Macedonian	Kazakh	Tetun	Georgian	Albanian
pysbd	**86.8**	**87.8**	**90.9**	**89.6**	82.0	94.0	92.1	**91.3**
nltk	80.7	35.6	85.6	82.9	82.0	85.5	91.8	88.0
stanza	80.3	83.4	88.9	84.0	**89.1**	**94.6**	**93.2**	76.1

Table 6: Sentence splitting performance (Accuracy %) per language. The best score for each language is highlighted in bold.

C OpenWHO performance per language

Model	mkd	kaz	kat	hye	aze	sqi	tet	som	sin	AVG
MADLAD-400 10B	58.37 / 3.25	47.29 / 4.59	15.81 / 14.09	37.27 / 5.25	40.54 / 6.24	54.97 / 3.87	44.29 / 7.35	48.13 / 6.44	39.48 / 6.04	42.73 / 6.22
NLLB-200 3.3B	50.39 / 3.18	42.94 / 3.64	38.27 / 4.22	39.19 / 4.20	45.23 / 4.32	57.50 / 3.09	– / –	47.05 / 4.40	39.69 / 3.40	45.03 / 3.81
NLLB-200 54B	56.17 / 2.94	56.55 / **3.15**	43.90 / 4.21	42.91 / 3.62	48.78 / 3.92	59.01 / 2.84	– / –	48.23 / 4.50	48.64 / 3.04	50.52 / 3.53
Gemma-3 27B	58.52 / 2.95	48.90 / 3.99	43.76 / 4.66	43.37 / 4.09	46.09 / 4.36	58.12 / 3.11	36.85 / 8.82	46.01 / 5.62	39.32 / 5.11	48.01 / 4.24
DeepSeek-V3 671B	59.07 / 2.98	50.39 / 3.69	46.27 / 3.78	47.41 / 3.39	47.84 / 3.73	59.02 / 2.89	47.64 / 7.12	43.36 / 6.18	41.70 / 4.68	49.38 / 3.92
Gemini 2.5 Flash	**62.83 / 2.65**	**57.41** / 3.17	**50.28 / 3.00**	**49.40 / 3.00**	**52.62 / 3.51**	**60.33 / 2.72**	**51.86 / 6.22**	**55.09 / 4.15**	**54.58 / 2.58**	**55.32 / 3.10**

Table 7: Overall performance (ChrF++ / MetricX) on the OpenWHO test set. LLM scores represent their optimal strategy (max in Table 8). The best score in each column is in **bold**.

Model	Strategy	mkd	kaz	kat	hye	aze	sqi	tet	som	sin	AVG	Δ
Gemini 2.5 Flash	Sentence level	57.8	54.9	46.2	46.1	48.6	57.0	46.8	52.2	50.9	51.2	–
	Sentence window	58.5	54.8	47.2	45.6	48.9	58.5	46.7	48.9	50.1	51.0	−0.2
	Sentence + doc context	58.6	54.9	45.4	46.4	50.2	57.8	47.3	51.3	51.4	51.5	+0.3
	Document level	62.5	**57.4**	49.0	49.1	51.6	60.0	51.5	55.0	**54.6**	54.5	+3.3
	Doc-level + self-correct	**62.8**	56.3	**50.3**	**49.4**	**52.6**	**60.3**	**51.9**	**55.1**	54.5	**54.8**	+3.6
DeepSeek-V3	Sentence level	57.0	49.2	42.6	44.3	45.2	57.0	43.5	41.2	40.7	46.7	–
	Sentence window	57.8	49.5	44.9	45.4	47.6	58.1	44.5	42.8	**41.7**	48.0	+1.3
	Sentence + doc context	58.3	49.7	44.2	44.6	46.6	57.9	43.3	42.1	40.2	47.4	+0.7
	Document level	**59.1**	**50.4**	**46.3**	45.0	**47.8**	**59.0**	**47.6**	42.9	40.7	**48.7**	+2.0
	Doc-level + self-correct	55.7	46.3	43.3	**47.4**	47.4	57.1	45.7	**43.4**	41.3	47.5	+0.8
Gemma-3 27B	Sentence level	58.1	48.3	**43.8**	**43.4**	45.8	**58.1**	35.3	**46.0**	38.6	**46.4**	–
	Sentence window	58.3	48.5	43.1	43.0	45.7	58.0	32.7	45.4	37.7	45.8	−0.6
	Sentence + doc context	56.6	47.5	41.0	40.2	43.1	56.8	32.6	44.4	38.0	44.5	−1.9
	Document level	**58.5**	**48.9**	40.9	42.5	**46.1**	57.5	36.5	45.2	**39.3**	46.2	−0.2
	Doc-level + self-correct	57.4	46.2	42.6	41.5	46.0	57.3	**36.9**	45.4	38.3	45.7	−0.6

Table 8: Effect of different context strategies on LLM performance on the OpenWHO test set (ChrF++). The ‘Δ‘ column shows the change relative to the ‘sentence level‘ baseline.

D WMT24++ performance per language

Model	tam	zul	bul	srp	swh	AVG
NLLB-200 3.3B	43.85 / 3.52	**63.35 / 3.17**	58.39 / 3.09	51.59 / 3.16	52.17 / 4.62	53.87 / 3.51
NLLB-200 54B	45.52 / 3.38	58.39 / 3.23	59.80 / 2.78	53.18 / 2.8	51.02 / 5.07	53.58 / 3.45
MADLAD-400 10B	40.68 / 3.98	38.24 / 5.69	59.40 / 2.88	47.38 / 5.38	46.02 / 7.1	46.34 / 5.01
Gemma-3 27B	45.14 / 2.99	43.95 / 5.93	59.55 / 2.50	52.36 / 2.62	52.56 / 3.83	50.71 / 3.61
DeepSeek-V3 671B	43.95 / 3.50	49.41 / 4.55	58.27 / 2.66	52.53 / 2.60	52.84 / 3.72	51.40 / 3.42
Gemini 2.5 Flash	**45.84 / 2.61**	54.77 / 3.25	**61.40 / 2.33**	**56.83 / 2.29**	**55.29 / 2.99**	**54.83 / 2.69**

Table 9: Overall performance (ChrF++ / MetricX) on the **WMT24++ news** test set. LLM scores represent their optimal context strategy (see Table 11).

Model	tam	zul	bul	srp	swh	AVG
NLLB-200 3.3B	30.74 / 7.85	46.51 / 4.99	47.29 / 4.80	42.17 / 5.44	45.16 / 6.29	42.37 / 5.87
NLLB-200 54B	30.91 / 7.99	46.55 / 4.92	48.63 / 4.57	44.27 / 4.93	44.64 / 6.74	43.00 / 5.83
MADLAD-400 10B	27.44 / 9.21	35.42 / 6.76	47.38 / 4.75	37.73 / 7.02	38.67 / 7.94	37.33 / 7.14
Gemma-3 27B	38.29 / 4.59	38.26 / 6.71	54.98 / 3.45	47.94 / 3.89	47.93 / 5.24	45.48 / 5.26
DeepSeek-V3 671B	38.10 / 4.99	44.33 / 5.64	53.70 / 3.55	49.58 / 3.61	48.69 / 5.08	46.88 / 4.57
Gemini 2.5 Flash	**39.47 / 3.94**	**50.99 / 4.30**	**57.60 / 3.30**	**53.21 / 3.18**	**52.03 / 4.09**	**50.66 / 3.76**

Table 10: Overall performance (ChrF++ / MetricX) on the **WMT24++ literary** test set. LLM scores represent their optimal context strategy (see Table 11).

Model	Strategy	tam	zul	bul	srp	swh	AVG
Gemini	*News*						
	Sent-level	**45.92** / 2.65	53.07 / 3.33	59.93 / 2.47	53.09 / 2.39	**55.90** / 3.03	53.58 / 2.77
	Doc-level	45.84 / **2.61**	**54.77 / 3.25**	**61.40 / 2.33**	**56.83 / 2.29**	55.29 / **2.99**	**54.83 / 2.69**
	Literary						
	Sent-level	34.41 / 5.71	44.34 / 5.12	49.80 / 4.41	44.91 / 4.66	48.00 / 4.82	44.29 / 4.94
	Doc-level	**39.47 / 3.94**	**50.99 / 4.30**	**57.60 / 3.30**	**53.21 / 3.18**	**52.03 / 4.09**	**50.66 / 3.76**
DeepSeek-V3	*News*						
	Sent-level	**43.95 / 3.50**	**49.41** / 4.60	**58.27** / 2.80	**52.53** / 2.73	**52.84** / 3.98	**51.40** / 3.52
	Doc-level	43.40 / 3.55	48.16 / **4.55**	56.96 / **2.66**	52.23 / **2.60**	52.14 / **3.72**	50.58 / **3.42**
	Literary						
	Sent-level	33.36 / 6.37	41.75 / 5.84	49.53 / 4.40	46.08 / 4.75	46.97 / 5.44	43.54 / 5.36
	Doc-level	**38.10 / 4.99**	**44.33 / 5.64**	**53.70 / 3.55**	**49.58 / 3.61**	**48.69 / 5.08**	**46.88 / 4.57**
Gemma-3 27B	*News*						
	Sent-level	**45.14 / 2.99**	**43.95 / 5.93**	59.55 / 2.62	52.36 / 2.69	**52.56 / 3.83**	**50.71 / 3.61**
	Doc-level	45.01 / 3.17	42.90 / 6.47	**60.28 / 2.50**	**52.43 / 2.62**	52.24 / 3.94	50.57 / 3.74
	Literary						
	Sent-level	34.53 / 5.84	**38.26 / 6.71**	50.29 / 4.27	42.66 / 4.84	46.21 / 5.33	42.39 / 5.40
	Doc-level	**38.29 / 4.59**	33.11 / 9.11	**54.98 / 3.45**	**47.94 / 3.89**	**47.93 / 5.24**	**44.45 / 5.26**

Table 11: Comparison of sentence-level and document-level strategies on the WMT24++ test set (ChrF++ / MetricX). The colored delta in the 'AVG' column shows the change relative to the 'sentence-level' baseline within each domain.

E AutoMQM results

Model	mkd	kaz	kat	hye	aze	sqi	som	sin	AVG
AutoMQM score (lower is better)									
NLLB-54B	-4.72	-3.35	-5.44	-4.34	-4.27	-3.11	-6.08	-4.54	-4.48
Gemini 2.5 Flash	**-2.80**	**-3.15**	**-3.12**	**-2.55**	**-3.01**	**-2.59**	**-4.88**	**-2.40**	**-3.06**
Difference in error counts (Gemini - NLLB)									
Error category	**mkd**	**kaz**	**kat**	**hye**	**aze**	**sqi**	**som**	**sin**	**AVG**
Accuracy									
Mistranslation	-4	-4	-33	-13	-5	-15	-25	-22	↓ **-15.1**
Overtranslation	-6	19	-7	13	-2	2	18	7	↑ **+5.5**
Undertranslation	0	-4	-9	0	2	-3	-7	-5	↓ **-3.3**
Addition	4	1	4	1	-5	-3	2	-4	→ **0.0**
Omission	5	13	1	3	1	4	-1	5	↑ **+3.9**
Untranslated	-10	0	-8	-1	-1	-4	-4	-5	↓ **-4.1**
Total Accuracy	**-11**	**25**	**-52**	**3**	**-10**	**-19**	**-17**	**-24**	↓ **-13.1**
Fluency									
Grammar	-1	-4	-8	-7	-1	5	-14	-6	↓ **-4.5**
Spelling	-7	-2	0	-10	-2	4	4	-81	↓ **-11.8**
Punctuation	-1	-15	-1	-9	-1	3	0	-5	↓ **-3.6**
Total Fluency	**-9**	**-21**	**-9**	**-26**	**-4**	**12**	**-10**	**-92**	↓ **-19.9**
Style									
Awkward	3	12	-11	-2	-7	11	10	0	↑ **+2.0**
Register	0	2	-2	-1	-1	-5	3	-2	↓ **-0.8**
Total Style	**3**	**14**	**-13**	**-3**	**-8**	**6**	**13**	**-2**	↑ **+1.3**
Terminology									
Inconsistent	-5	1	0	1	3	-1	0	4	↑ **+0.4**
Wrong	-8	1	-10	-7	-7	0	6	-12	↓ **-4.6**
Total Terminology	**-13**	**2**	**-10**	**-6**	**-4**	**-1**	**6**	**-8**	↓ **-4.3**
Non-translation	**-4**	**0**	**-10**	**-3**	**0**	**-1**	**-4**	**-1**	↓ **-2.9**

Table 12: AutoMQM analysis comparing NLLB-54B and Gemini 2.5 Flash (sentence-level) on the OpenWHO test set. **Top**: Overall MQM scores (higher is better). Gemini consistently outperforms NLLB. **Bottom**: Difference in error counts (Gemini errors minus NLLB errors) per category. Negative values indicate Gemini made fewer errors for that category. Gemini outputs less major errors like mistranslations and incorrect terminology, at the cost of a slight increase in over-translation and omissions.

F Prompts

System: Translate from English to [target lang name]. Give only the translation, and no extra commentary, or chattiness. Wrap the translated sentence in <result></result> tags.

User: <text to translate>She lives in Boston.</text to translate>

Assistant: <result>[Google Translate of "She lives in Boston." into target lang]</result>

User: <text to translate>[sentence to translate]</text to translate>

Prompt used for **Sentence level** translation. We ask the model to wrap the translation in <result> tags to avoid model commentary interfering with translation accuracy measurement.

System: Using the provided context, translate the "Sentence to translate" from English to [target lang name]. Give only the sentence translation, and no extra commentary, or chattiness. Wrap the translated sentence in <result></result> tags.

User: <context>
Her name is Mary. She lives in Boston. She is a doctor.
</context>
Sentence to translate:
She lives in Boston.

Assistant: <result>[Google Translate of "She lives in Boston." into target lang]</result>

User: <context>
[preceding sentence][sentence to translate][next sentence]
</context>
Sentence to translate:
[sentence to translate]

Prompt used for **Sentence window** translation.

System: Using the provided context, translate the "Sentence to translate" from English to [lang name]. Give only the sentence translation, and no extra commentary, or chattiness.

User: <context>
Her name is Mary. She lives in Boston. She is a doctor.
</context>
Sentence to translate:
She lives in Boston.

Assistant: <result>[Google Translate of "She lives in Boston." into target lang]</result>

User: <context>
[whole document for the sentence]
</context>
Sentence to translate:
[sentence to translate]

Prompt used for **Sentence + doc context** translation.

System: Translate from English to [lang name]. Give only the translation, and no extra commentary, or chattiness. Use the same formatting as the source text to translate, with one sentence per line. Enclose your translation in <result></result> tags.

User: <text to translate>
Her name is Mary.
She lives in Boston.
She is a doctor.
</text to translate>

Assistant: <result>
[Google Translate of "Her name is Mary." into target lang]
[Google Translate of "She lives in Boston." into target lang]
[Google Translate of "She is a doctor." into target lang]
</result>

User: <text to translate>
[document sentence 1]
[document sentence 2]
...
</text to translate>

Prompt used for **Document level** translation.

System: Translate from English to [lang name]. Give only the translation, and no extra commentary, or chattiness. Use the same formatting as the source text to translate, with one sentence per line. Enclose your translation in <result></result> tags.

User: <text to translate>
Her name is Mary.
She lives in Boston.
She is a doctor.
</text to translate>

Assistant: <result>
[Google Translate of "Her name is Mary." into target lang]
[Google Translate of "She lives in Boston." into target lang]
[Google Translate of "She is a doctor." into target lang]
</result>

User: <text to translate>
[document sentence 1]
[document sentence 2]
...
</text to translate>
Assistant: [assistant response from above]
User: Please translate again for a better version. Be particularly mindful of using the right script and tone, of adapting to context, and of translating each sentence faithfully.
<text to translate>[same as above]</text to translate>]

Prompt used for **Doc-level + self-correct** translation.

Factors Affecting Translation Quality in In-context Learning for Multilingual Medical Domain

Jonathan Mutal, Raphael Rubino, Pierrette Bouillon
TIM/FTI, University of Geneva
1205 Geneva, Switzerland
firstname.surname@unige.ch

Abstract

Multilingual machine translation in the medical domain presents critical challenges due to limited parallel data, domain-specific terminology, and the high stakes associated with translation accuracy. In this paper, we explore the potential of in-context learning (ICL) with general-purpose large language models (LLMs) as an alternative to fine-tuning. Focusing on the medical domain and low-resource languages, we evaluate an instruction-tuned LLM on a translation task across 16 languages. We address four research questions centered on prompt design, examining the impact of the number of examples, the domain and register of examples, and the example selection strategy. Our results show that prompting with one to three examples from the same register and domain as the test input leads to the largest improvements in translation quality, as measured by automatic metrics, while translation quality gains plateau with an increased number of examples. Furthermore, we find that example selection methods – lexical and embedding based – do not yield significant benefits over random selection if the register of selected examples does not match that of the test input.

1 Introduction

Multilingual communication in clinical settings is often hindered by the lack of quality translation tools for low-resource languages (Zappatore and Ruggieri, 2024). Building machine translation (MT) systems in the medical domain is challenging: parallel corpora is scarce and mistakes can lead to disastrous outcome (Chan et al., 2024). This challenge is intensified when translating into or from low-resource languages (Phan et al., 2023). Traditional neural MT models require supervised training on domain-specific data, which is not feasible for many low-resource language pairs. On the other hand, large language models (LLMs) possess broad world knowledge through pre-training and can be instructed to perform various tasks. Yet, general-purpose LLMs, when translating only with instructions (in zero-shot setting), often fail to produce adequate translations for specialized domains (Neves et al., 2024; Hu et al., 2024).

In-Context Learning (ICL) offers a way to guide LLMs at inference time by providing a few input-output pairs examples as part of the prompt (Brown et al., 2020). Unlike fine-tuning, ICL does not update model parameters; instead, the model learns from examples on the fly. This approach has gained popularity for low-resource scenarios (Zebaze et al., 2025), since only a handful of examples (as few as 1–5) can significantly improve performances on a given task. Prior studies have explored various strategies to optimize ICL for MT (Vilar et al., 2023). For instance, selection of examples that are similar to the test input yields better translation output. Similarity can be defined lexically (word overlap) or semantically (vector distance), and there is ongoing debate on which is more effective (Zebaze et al., 2025). Recent work has also examined the impact of the number of examples: some found that using up to 5 examples is beneficial (Zhu et al., 2024a), while others observed improvements up to 8 examples before performance plateau (Zhu et al., 2024b). Additionally, domain match is believed to be important: examples from the same domain as the task can guide the model's lexical and stylistic choices (Agrawal et al., 2023; Aycock and Bawden, 2024). However, to the best of our knowledge, ICL for MT in the medical domain is yet to be explored.

Our work aims to provide an evaluation of in-context learning for low-resource medical text translation. In particular, our goal is to quantify the impact of three key factors that may influence translation quality: the number of in-context examples, the register of those examples, and the strategy used to select them. We evaluate how each factor affects translation quality using two automatic metrics, namely ChrF++ (Popović, 2015)

Proceedings of the Tenth Conference on Machine Translation, pages 161–179
November 8-9, 2025

and COMET (Rei et al., 2020). More precisely, we seek answers to the following research questions (RQ1, RQ2, RQ3 and RQ4):

RQ1: Effect of Number of Examples – Does increasing the number of in-context examples improve translation quality?

RQ2: Effect of Register – Do examples with matching registers yield better translations than mismatched ones?

RQ3: Effect of Selection Strategy – Does semantic similarity (content-based) versus lexical similarity (form-based) selection of examples impact translation quality?

RQ4: Effect of Linguistic Characteristics - Which linguistic characteristics of the in-context examples (corpus- and prompt-level) most strongly influence translation quality?

Our contributions include: (1) an extensive empirical evaluation on 16 diverse languages (covering African, European and Asian languages and dialects) in a medical setting (Section 4), (2) a statistical analysis of the different factors, individually and in combination, that contribute to translation quality (3) a linguistic analysis of how corpus and prompts characteristics affect translation quality measured by automatic metrics (Section 5).

2 Related Work

Low-resource MT A low-resource language is typically defined as a language for which limited annotated data, such as parallel corpora or monolingual text, is available for training data-driven NLP models. The lack of resources may apply to text quantity, domain coverage, or availability of evaluation benchmarks (Joshi et al., 2020). In MT research, a low-resource language pair refers to a translation direction where parallel corpora are insufficient to train reliable MT systems. This limitation may reflect the absolute size (e.g., <1M sentence pairs), the domain (e.g., biomedical), or the bilingual coverage (Koehn and Knowles, 2017). In (bio)medical machine translation, even high-resource languages can become low-resource in-domain, due to the scarcity of domain-specific aligned corpora (e.g., medical records, Cochrane reviews, or medical dialogues) (Neves et al., 2024).

ICL for MT Early work on prompting LLMs for MT showed that models like GPT-3 can perform translation tasks without fine-tuning, particularly for high-resource languages (Brown et al., 2020). However, for low-resource languages and specialized domains, zero-shot performance is often weak, motivating research into few-shot prompting techniques (Hendy et al., 2023). One factor shown to influence translation quality is the number of examples. For instance, Peng et al. (2023) found improvements up to about five examples, while Zhu et al. (2024c) reported gains up to eight examples before saturation (translation quality scores plateau). These differences suggest that task- or model-specific characteristics can affect translation quality. This research direction allows us to answer RQ1.

Few-shot selection The example selection strategy for ICL is another impacting factor for MT, which has produced mixed results. Vilar et al. (2023) did not observe significant differences between random examples and lexically similar examples in some setups. In contrast, Zebaze et al. (2025) found that a semantic or lexical selection of examples based on the similarity with the source text to be translated improves translation quality. Moslem et al. (2023) proposed a more fine-grained approach, identifying which source words contribute most to guide example selection. More recently, Zebaze et al. (2025) showed that a small number of similar examples can yield large gains in translation quality for low-resource languages, even if the impact is limited for high-resource pairs where the LLM is already strong. We contribute to this debate by comparing form-based and meaning-based retrieval (answering RQ3), using BM25 (Robertson and Zaragoza, 2009) and LASER (Artetxe and Schwenk, 2019), respectively. This comparison has not been extensively explored in a multilingual medical setting.

Domain, Register and MT Domain is another influencing factor in MT quality. Farajian et al. (2017) observed that domain-matching data between training and testing leads to MT improvement, which motivates our experiments on few-shot selection from various sample pools, including in- and out-of-domain corpora. Following these ideas, Agrawal et al. (2023); Sia and Duh (2023); Aycock and Bawden (2024) showed that using ICL for MT, using in-domain data (e.g., medical or legal) as examples helps the model to produce appropriate terminology and style. In this work, we examine not only the domain but also the register (Lecorvé et al., 2023) of the texts (RQ2), as both can influence translation quality – an aspect that has received little attention in prior work. While

domain refers to the subject matter of a text, such as medicine, law, or education, register describes how language is used in a specific situation within that domain, shaped by factors like the relationship between speakers, the communication channel, and the purpose of the interaction. For example, both a doctor–patient dialogue and a medical research paper belong to the medical domain while being in different registers and thus exhibit different styles, choices in vocabulary, and overall communicative intent.

3 Methodology

3.1 Test Data

Our evaluation spans 16 languages: Albanian, Modern Standard Arabic, Moroccan Arabic, Tunisian Arabic, Dari, Farsi (Persian), Russian, Romanian, Ukrainian, English, French, Spanish, German, Polish, Czech and Tigrinya. These include low-resource languages/dialects (e.g. Moroccan, Tunisian and Tigrinya) as well as higher-resource ones (French, Spanish). We consider translation between all pairs and translation directions among these languages.

We evaluate various prompt engineering settings on three test sets in the medical domain that differ in register. These test sets are n-way parallel–each sentence translated into multiple languages–thus allowing us to assess: i) cross-linguistic variations, ii) differences in style and iii) communicative purposes within medical texts.

Cochrane is an internationally recognized source of evidence-based clinical research, providing reviews that synthesize medical studies to inform clinical practice[1]. The language used in Cochrane documents is formal, technical, and structured, making it representative of technical biomedical content. The content of this article is aimed at medical professionals, particularly researchers in the medical field.

NHS24 consists in publicly available health articles from Scotland's national telehealth service [2]. These articles are designed for the general public and provide accessible medical information, symptom explanations, and healthcare guidance. The language used in this article aims to be understandable, non-technical, and oriented toward patient comprehension, distinguishing it from more technical registers such as Cochrane. This corpus represents a patient-facing, health communication register, in a scenario where translation clarity and simplicity are critical.

Medical Dialogues is a set of medical question-and-instruction sentences from Bouillon et al. (2021). This corpus has never been released publicly, thus constitute an annotated *no-leakage* evaluation set never seen by LLMs. Sentences in this corpus are characterised by short, directive, and information-seeking utterances typical of clinician–patient interactions (e.g., asking about symptoms, giving instructions for treatment, etc.). The sentences were translated from French.[3]

An important characteristic of our medical dialogue data is that the translators were instructed to generate target-oriented translation that read as if originally written in the target language. They were also asked to take the communicative context and audience into account (e.g., patient vs. clinician) and allowed freedom of reformulation rather than adhering to the source structure. As a result, the dataset avoids many of the typical artifacts of translationese – such as literal lexical choices, unnatural word order, or oversimplification – while still maintaining terminological accuracy (Gerlach et al., 2018). An example of the translations is shown in Table 2. For the other datasets, we acknowledge the potential influence of translationese, but because both training and evaluation rely on the same language pairs (including artificial test pairs), the artifacts introduced by translationese is expected to be consistent across sets and therefore less likely to distort automatic scores (Ni et al., 2022).

3.2 Register and Domain Selection

To assess the effect of test and n-shot source mismatch, various datasets are used as sources for few-shot sampling: datasets described in Section 3.1, as well as a general-domain corpus: FLORES+ (Team et al., 2022). In our experiments, we control for all other factors (model, language pair, test sentence, prompt length, and evaluation metric) and vary only the source of the in-context examples, allow-

[1] An example can be found at https://pmc.ncbi.nlm.nih.gov/articles/PMC7045447/

[2] For instance, https://www.nhsinform.scot/healthy-living/preventing-falls/falls-and-dementia/

[3] The dataset is available in https://huggingface.co/datasets/jonathanmutal/Medical-Questionnaire-Multilingual-Translation

ing us to compare matched vs. mismatched domain and register. We follow the standard dataset splits for Cochrane and NHS24 (Haddow, 2015) and for FLORES+ (Team et al., 2022). For the Medical Dialogues set, we randomly extracted 1,000 segments. The number of n-way parallel segments for each dataset is shown in Table1.

Dataset	Split	#Sentences
Cochrane	Train	759
	Test	672
NHS24	Train	1200
	Test	1257
Medical Dialogues	Train	8511
	Test	1000
FLORES+	Train	997

Table 1: Number of sentences for each dataset and split.

3.3 In-Context Example Selection

To address RQ3, we consider three example selection strategies:

Random: We randomly sample n examples from the available pool of parallel data. This acts as a baseline and helps quantify variance. To ensure fairness, the same set of random examples (for a given n) is used across different languages when evaluating the number of examples. This way, any observed differences are not due to content variations across languages.

Lexical Similarity (BM25): We retrieve examples that are lexically similar to the input, using the BM25 (Robertson and Zaragoza, 2009) ranking function. BM25 scores examples based on overlapping words (with term-frequency and length normalization). This method prioritizes examples containing similar medical terms or phrases, reinforcing consistent terminology.

Semantic Similarity (LASER): We use multilingual LASER embeddings (Artetxe and Schwenk, 2019) to find examples with high cosine similarity to the input sentence. This can retrieve examples that are paraphrases or semantically related, even if they do not share keywords. The goal is to help the model generalize to similar meanings expressed using various surface forms.

For both BM25 and LASER, we retrieve the top n examples from each dataset individually. We also follow a specific ordering when placing examples in the prompt: we sort the retrieved examples by descending similarity to the input (most similar last, closest to the input). This ordering, suggested by prior work (Chitale et al., 2024), may maximize the utility of the demonstration closest to the test query.

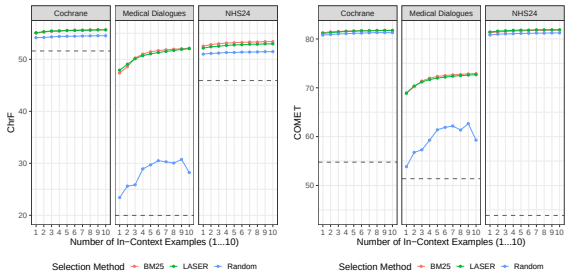


Figure 1: Effect of the number of in-context examples on ChrF scores across datasets and selection methods. ChrF performance is plotted for three datasets – Cochrane, Medical Dialogues, and NHS24– using three selection strategies: BM25 (red), LASER (green), and Random (blue). Scores are averaged across all languages. The dashed horizontal lines represent zero-shot performance.

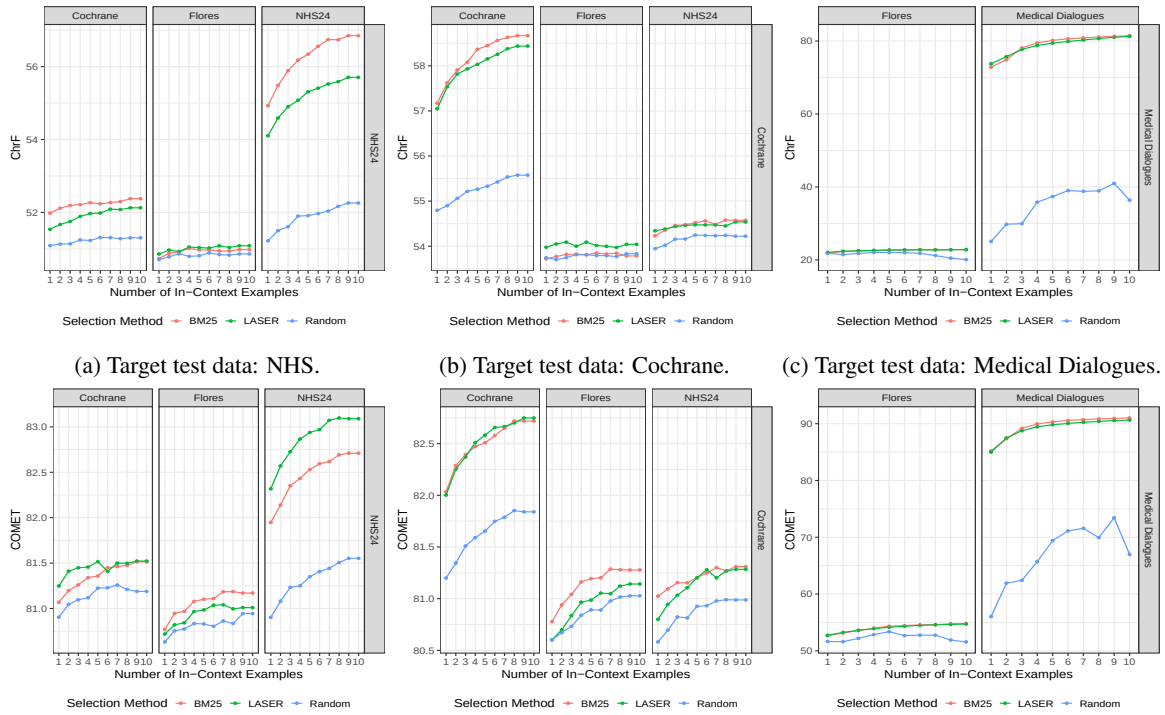
3.4 Experimental Settings

We use Mistral-7B-Instruct[4] with a fixed JSON-based prompt format to ensure systematic outputs and isolate the effect of example content on translation quality (cf. Appendix B for more details). We vary the number of in-context examples n from 0 (zero-shot, only instruction) up to 10 to address RQ1. Each configuration (defined by the number of examples n, the selection method, and the domain of the selected examples) is applied to all test sentences. For stochastic settings (random selection), we repeat each test 30 times with different random seeds and average the results. This yields robust estimates of performance by smoothing the results obtained with random sampling and allows for significance testing.

We evaluate translation quality with two automatic metrics:

- ChrF++: Character n-gram F-score, which correlates well with adequacy especially for morphologically rich languages (Popović, 2015).
- COMET: A learned metric that predicts human judgment scores using multilingual em-

[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3



(a) Target test data: NHS. (b) Target test data: Cochrane. (c) Target test data: Medical Dialogues.

(d) Target test data: NHS. (e) Target test data: Cochrane. (f) Target test data: Medical Dialogues.

Figure 2: Effect of example domain on translation quality across selection methods. ChrF and COMET scores are shown for in-context examples drawn from different domains using three selection methods (BM25, LASER, RANDOM).

beddings; effective for capturing semantic adequacy (Rei et al., 2020).

All metrics are computed against reference translations. We conduct significance testing Factorial Analysis of Variance (Factorial ANOVA, Ross and Willson, 2017) to understand the effect of each factor (number of examples, register and selection strategy) on the translation quality, and also understand the effect of multiple factors on translation quality. This analysis allows us to understand, for example, the effect of the number of examples and selection strategy on the translation quality. We used eta squared (η^2) test (Adams and Conway, 2014) to quantify the effect size in analysis of variance and determine which factor has the largest effect on translation quality according to automatic metrics.

4 Effect of Factors in Translation Quality

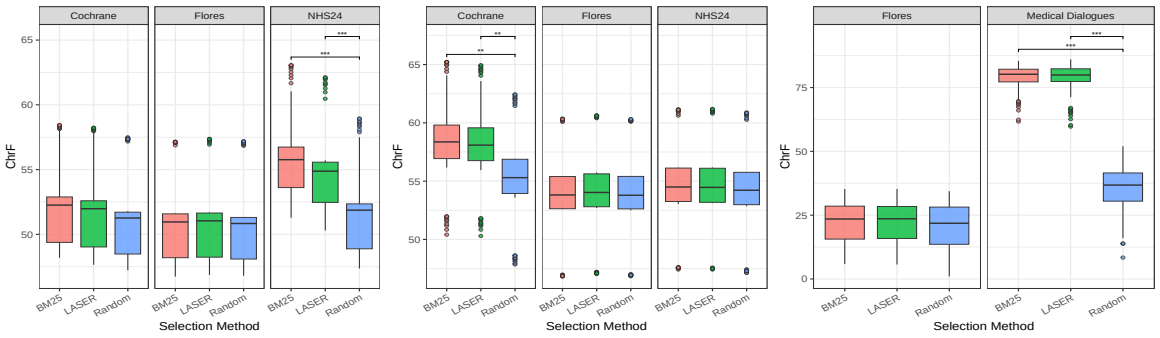
In this section, we describe the results collected during our experiments. We divide the results following the RQ1, RQ2 and RQ3 from Section 1 before comparing the effects of all factors on translation quality.

4.1 Results

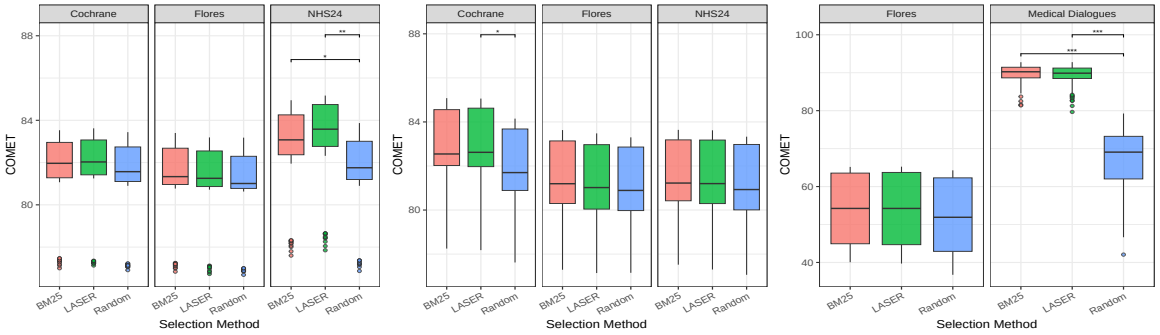
Number of Examples: Figure 1 illustrates the effect of the number of examples on translation quality. We observe that increasing the number of in-context examples improves translation quality for all test sets, but differences are observed in terms of n-shot configuration.

With no translation examples in the prompt (0-shot, the dashed line in Figure 1), the LLM reaches the lowest scores on the medical dialogues test set. With just a single example, automatic scores more than doubled (from 20.01 to 48.20 ChrF using BM25). For the best selection method on this test set (medical dialogues), there are diminishing returns above 4 to 5-shot configuration.

For Cochrane and NHS24, the difference between 0-shot and n-shot is smaller based on ChrF (51.25 vs. 55.21 for Cochrane, 45.23 vs. 53.25 for NHS24). For these particular test sets, most gains are obtained with 2 to 3 examples, followed by a plateau with no significant improvement when increasing the number of shots. This may be due to the fact that the Cochrane and NHS24 datasets are publicly available and may have been seen by the LLM during pre-training, whereas this is not



(a) Target test data: NHS. (b) Target test data: Cochrane. (c) Target test data: Medical Dialogues.



(d) Target test data: NHS. (e) Target test data: Cochrane. (f) Target test data: Medical Dialogues.

Figure 3: Effect of selection method on translation quality across different test datasets. ChrF and COMET scores are shown for in-context examples drawn from different selection methods (BM25, LASER, RANDOM) on the different domain examples. Statistical significance among results is indicated by *** when $p < 0.001$, ** when $p < 0.01$ and * when $p < 0.05$.

the case for Medical Dialogues.

To answer RQ1, these results show that increasing the number of examples has an effect on translation scores measured by automatic metrics. However, a plateau is quickly reached on all test sets (max. with a 5-shot prompt), and adding more examples does not lead to significant improvements. This is observed when examples are retrieved based on their similarity with the source using BM25 and LASER.

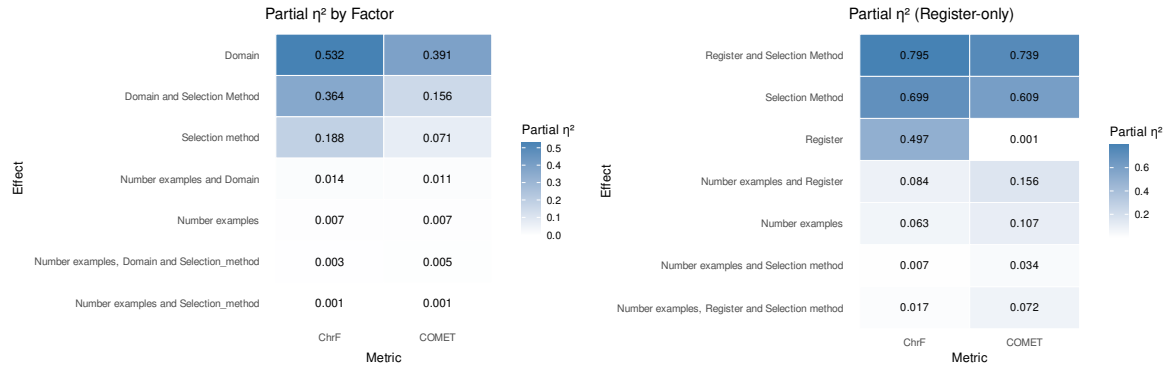
Source of Examples: Figure 2 shows translation results on our test sets when various sample pools are used to build the prompt.

Matching-domain exemplars (n-shot and test sentences sampled from the same corpus) result in significant score gains for all selection methods. The figures show that sampling from FLORES+ consistently underperforms compared to sampling from in-domain datasets. We also observe that random sampling from in-domain data outperforms a selection strategy using an out-of-domain dataset, showing that the domain of the data have a strong effect on translation quality.

Additionally, results show that matching registers yields the highest scores. However, unlike domain mismatch, we found that using a non-random selection method with Cochrane on the NHS24 test set yields higher ChrF scores compared to randomly sampling from the same register (sampling from NHS24), although the difference is not statistically significant according to COMET. This may be due to the content of the NHS24 corpus, written for patients and thus using less technical vocabulary compared to Cochrane.

To answer RQ2, the results show that register and domain have a significant effect on translation quality.

Example Selection Method: We compare lexical vs. semantic retrieval (BM25 vs. LASER, respectively) against random examples selection to address RQ3. Figure 3 illustrates the box plot for the different test data with the n-shot retrieval methods. Overall, BM25 and LASER yield nearly identically scores on automatic metrics. BM25 had a slight higher automatic scores (but not statistically significant). We found that when using BM25 for lower-resource languages seems to be more beneficial (we refer to Figure 10 for these



(a) Partial η^2 for different domains and registers.

(b) Partial η^2 for the same domain and different registers.

Figure 4: Heatmap of partial η^2 values indicating the percentage of variance explained by each factor and interaction in the model. Darker shades represent greater effect on translation quality. Register and Selection method show the highest effects, while all interactions involving Number examples contribute minimally. Note that partial eta-squared values are not additive and do not sum to 100%.

results).

BM25 and LASER achieve higher translation quality, as measured by the automatic metrics, compared to random sampling when the domain and register match those of the test data. However, when using data from a different domain or register, the selection methods do not yield significant improvements compared to random sampling. This provides further evidence of n-shot domain and register impact on translation quality, adding support to the findings for RQ2. We can therefore answer RQ3: there is an effect of selection strategy when the examples match the register. Otherwise, there is no statistical difference between BM24, LASER and random selection.

Comparing Effects: We conducted a factorial ANOVA to quantify the contribution of each factor to the variance in translation quality, as measured by ChrF and COMET, illustrated by Figure 4a.

Using common benchmarks for partial η^2 ($\sim$ .01 small, $\sim$ .06 medium, $\sim$ .14 large)[5], this analysis reveals that register and domain shows the largest effect ($\eta^2 = 0.53$), suggesting the highest variance in translation quality is associated with whether the train and test are sampled from the same dataset. The selection method ($\eta^2 = 0.18$) has also a large impact on translation quality. However, the interaction between the selection method and matching dataset effects on translation quality is higher (η^2 = .36), indicating that the selection method has a larger impact only when accompanied by matching n-shot register and domain.

[5] https://resources.nu.edu/statsresources/eta

On the other hand, increasing the number of examples in the prompt does not seem to have a strong impact ($\eta^2 = .0028$) compared to the other factors, i.e. domain and register. This supports the importance of the data source for n-shot selection. Increasing the number of examples provided as prompt to the LLM shows small additional variance on translation quality (above 1-shot).

When measuring translation quality variance within the medical domain, the main drivers are the selection method ($\eta^2 = 0.699$) and its interaction with the register ($\eta^2 = 0.795$). Number of examples is modest ($\eta^2 = 0.063$), and other interactions are small. Within the same domain, the choice of example selection strategy have a strong influence on both ChrF and COMET scores ($\eta^2 = 0.795$ and $\eta^2 = 0.739$ respectively), with its impact varying across registers. Figure 4b illustrates this values in a heatmap figure.

These results confirm our findings of the most important factor of translation quality: register and domain of the examples. To understand the causes, we carry out a lingistic evaluation in the following section.

5 Linguistic Analysis

Based on the previous results, which showed that the domain and register of the examples have the strongest impact on translation quality, we conduct a linguistic evaluation to determine which linguistic characteristics from the examples have the most influence on translation quality (RQ4). Because our goal is to assess how domain and register alignment shape translation quality, we select three corpus-

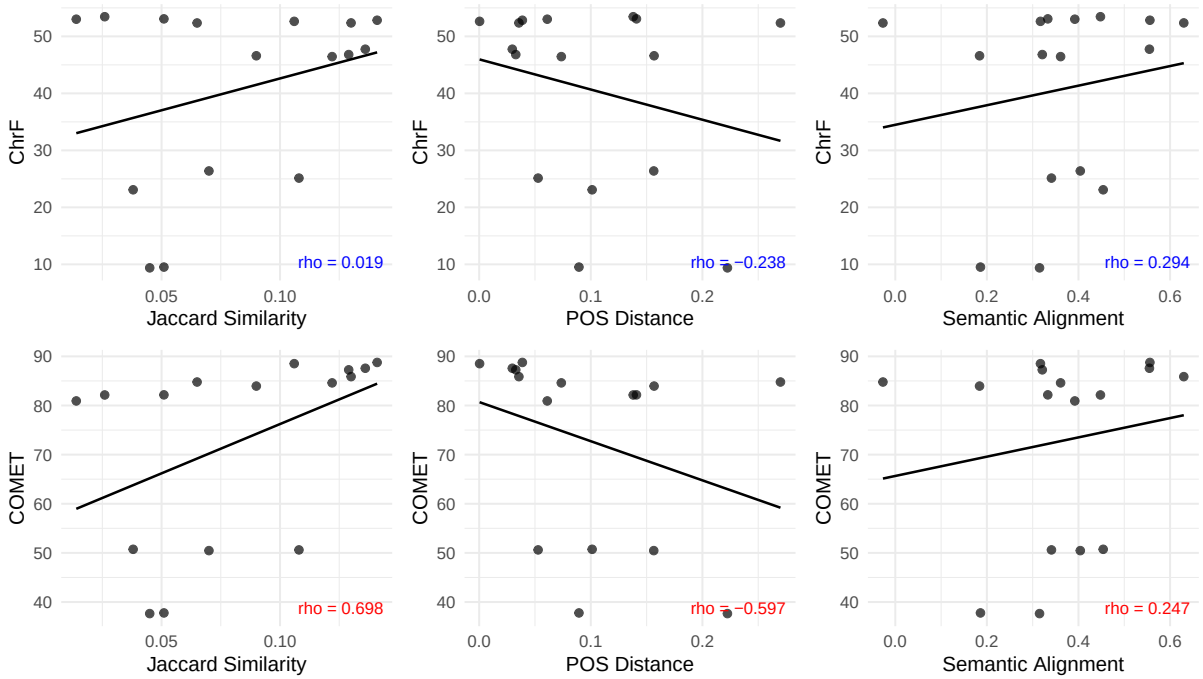


Figure 5: Scatterplots showing the relationship between corpus-level similarity metrics and average translation quality (ChrF and COMET). Each plot corresponds to one of the following corpus-level measurements: **Jaccard Similarity** (lexical overlap), **POS Distance** (cosine distance between part-of-speech distributions), and **Semantic Alignment** (average of sentence cosine similarity between corpus). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

level metrics that capture complementary aspects of similarity in texts:

- **Lexical overlap** (using (Jaccard similarity Jaccard, 1901) for corpus-level analysis and token overlap for prompt-level analysis): it measures vocabulary overlap across examples to capture whether they share key medical terms, and thus belong to the same domain and register. High overlap indicates coherence in subject matter.

- **Structural similarity** (cosine difference of part-of-speech distributions) (Liu et al., 2021): Part-of-speech distributions reflect grammatical choices; their cosine difference approximates whether examples adopt similar interpersonal stances and modes of medical communication.

- **Semantic alignment** (cosine similarity of sentence embeddings): Compares sentence embeddings to assess whether the overall meaning of the examples aligns. We used a different sentence embedding model from that used in the selection method, specifically the one proposed by (Reimers and Gurevych, 2019).

To perform this analysis, we included additional corpora for n-shot sampling to provide more data for corpus-level analysis: translations of documents related to public health and disease prevention across different languages within the European Union (ECDC, European Centre for Disease Prevention and Control, Greer, 2012); diverse datasets created during the COVID-19 period, including a set of Wikipedia documents related to health (Wikipedia Health); a database containing European Union law and other public documents generated during COVID-19[6]; and TICO-19 for non-European languages (Anastasopoulos et al., 2020). Finally, we included a general-domain corpus, Tatoeba[7]. While several of these texts share the same domain (medical), they differ in register, ranging from policy documents (EU public documents) to encyclopedic health texts and public health advisories. All corpora were extracted using the OPUS platform (Tiedemann, 2012).

We first examine corpus-level characteristics to understand how to select sample pools for ICL and to determine which aspects of domain and register are the most impactful when choosing a sample

[6]Extracted from https://elrc-share.eu/
[7]https://tatoeba.org/en/

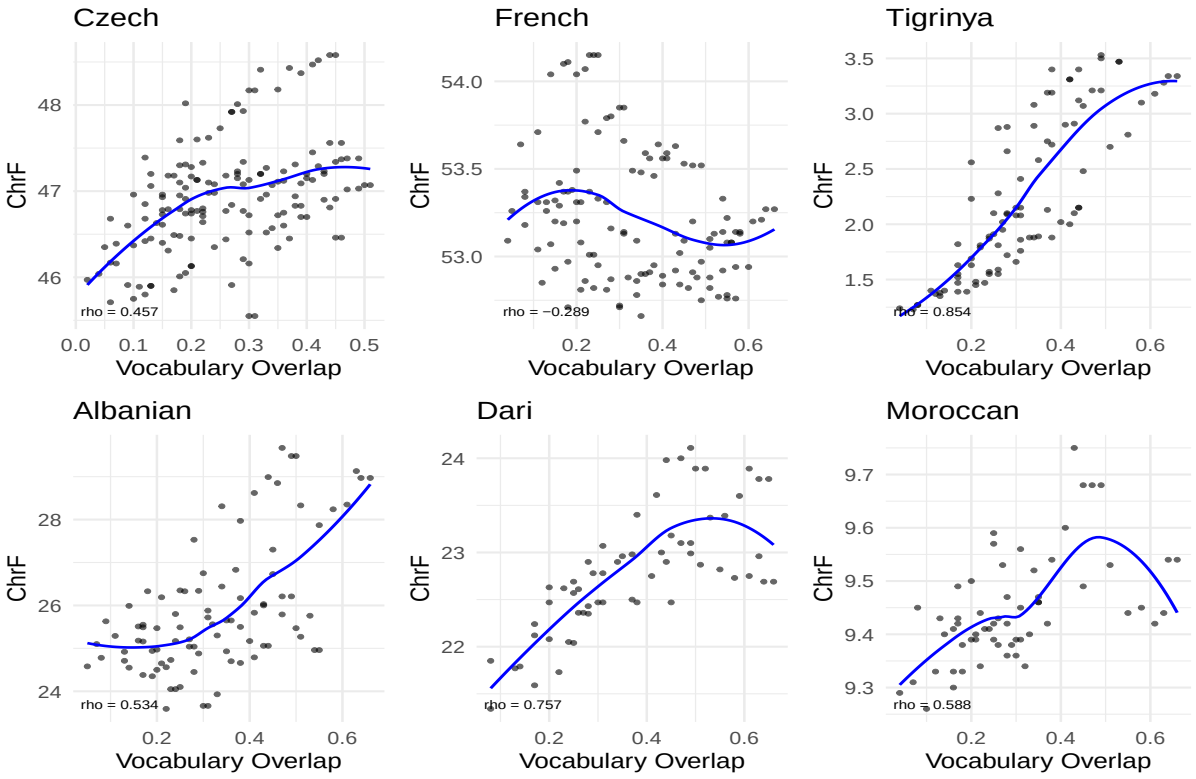


Figure 6: Scatterplots showing the relationship between prompt-level similarity metrics and average translation quality. Each plot corresponds to one of the following prompt-level measurement for **Vocabulary Overlap** (lexical overlap). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

pool. We then analysed prompt-level examples[8] to evaluate the effect of their linguistic characteristics on translation quality. To identify which aspects of register matter most, we examined the relationship between linguistic features and evaluation scores, checking how strongly each feature–such as vocabulary overlap, grammatical similarity, and semantic similarity–correlated with ChrF and COMET using Spearman rank correlation.

We selected language pairs in which POS tagging and embedding-based metrics for the source language were supported and for which the neccesary corpora were available. Specifically, we used English paired with six languages: two high-resource languages (French and Czech) and four low-resource languages (Tigrinya, Albanian, Moroccan Arabic, and Dari), selected based on the availability of medical corpora.[9]

[8]When multiple examples were provided for in-context learning, we calculated the mean; maximum scores were also tested but showed lower correlation with translation quality according to Pearson's ρ.

[9]According to the OPUS platform, English–Tigrinya has 6,142 sentences in the COVID medical domain; English–Dari has 3,071; English–Albanian has 389; and English–Moroccan Arabic has none.

In the next section, we describe the results collected during our experiences to answer RQ4, divided by corpus and prompt levels of analysis.

5.1 Results

Corpus-level Analysis At the corpus level, the similarity measures (lexical, syntactic, semantic) are the same for all sentences within a given sampling and test dataset, as they are calculated over the full example set, while translation quality varies between prompts. To avoid inflating the number of independent observations, we averaged the translation quality scores for all sentences in the same sampling and test dataset configuration and used these aggregated values in the analysis. This ensures that each configuration is counted once, and the results reflect real differences between configurations rather than repetition of identical feature values.

Figure 5 illustrates the corpus-level analysis between the translation automatic scores and the corpus-level metrics for lexical overlap. The results show that lexical overlap strongly correlates with COMET ($\rho = 0.698$), but not with ChrF, suggest-

ing that COMET is more sensitive to lexical content alignment. POS distance shows negative correlations with both metrics, especially with COMET, indicating that structural divergence between examples and test sets degrades ICL performance. Semantic alignment correlates moderately with both ChrF ($\rho = 0.294$) and COMET ($\rho = 0.247$), confirming that semantically coherent prompts are beneficial, although their predictive power is lower than the lexical alignment for COMET.

Prompt-level Analysis To assess the effect of samples register and domain from the prompt on translation quality, we first calculated the Spearman's ρ correlation between the translation quality scores and the linguistic features between the samples and the input sentence – vocabulary overlap, grammatical similarity, and semantic similarity.

Figures 6, 11, 12 show that translation quality–measured by ChrF and COMET–correlates with lexical, syntactic, and semantic similarity between the input and the selected examples. Spearman correlations indicate that low-resource languages such as Tigrinya, Dari, Moroccan Arabic, and Albanian exhibit the strongest correlations across all three similarity types, while higher-resource languages display more selective patterns. The ANOVA analysis, which includes the number of examples and the selection method as fixed effects, confirm these trends, with semantic similarity often producing the largest effect in translation quality for low-resource languages, and syntactic similarity dominating in Czech. η^2 analysis further reveals the unique contribution of each feature: semantic similarity explains the largest share of variance in most low-resource languages (e.g., 4–10% in Albanian and Moroccan Arabic), whereas in French lexical similarity accounts for 13–15% and in Czech syntactic similarity explains up to 21.7% of variance in ChrF. Together, these results show that the relative importance of lexical, syntactic, and semantic alignment is language-dependent.[10]

6 Conclusions

This study shows that, in multilingual medical machine translation, the domain and register of in-context examples are the most influential factors affecting translation quality. Partial η^2 analysis confirms that aligning the n-shot register and domain with the test input yields substantially greater improvements than increasing the number of examples. In practice, a small, well-chosen set of domain-relevant shots often yields higher translation quality scores than a larger set of examples sampled from other domains or registers.

Sentence-level analysis of lexical, syntactic, and semantic similarity confirms that the most predictive features vary by language. In low-resource language pairs, all three similarity types correlate strongly with translation quality, while in higher-resource languages pairs such as English to Czech and French, syntactic and semantic similarity dominate. Semantic similarity is the most consistent predictor across languages.

These results suggest that prompt engineering for ICL should prioritise register and domain alignment, and adapt exemplar selection criteria to the characteristics of the language pair rather than applying the same similarity heuristics.

Acknowledgements

This work is part of the PROPICTO project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005). We would also like to thank the three reviewers for their careful suggestions, which helped improve this work.

Limitations

This study has several limitations. First, all experiments were conducted using a single LLM, which constrains the generalisability of the findings to other model families, training paradigms and sizes. We hypothesize that larger models could reach better translation quality, which we leave for future work. Second, the linguistic similarity features—lexical, syntactic, and semantic—were computed using specific operationalisations (e.g., Jaccard similarity, POS distribution cosine distance, sentence embedding cosine similarity). They represent only one way of quantifying similarity, and alternative feature definitions or embeddings might yield different rankings of predictive importance. Moreover, corpus-level features were constant within each configuration, which required aggregation to avoid artificially statistical significance; this design limits the granularity of the corpus-level analysis. The linguistic evaluation was limited to a fixed set of high- and low-resource languages in the medical domain, meaning that

[10] Type–token ratio was negatively correlated with both metrics in nearly all languages, suggesting that higher lexical diversity in prompts tends to reduce translation quality.

results may not generalise to other languages. Finally, while ChrF and COMET provide complementary perspectives on translation quality, incorporating human evaluation for adequacy and fluency would strengthen the validity of the results. Furthermore, the evidence gathered in this work provides practical insights into the factors influencing translation quality as measured by automatic metrics. However, these findings do not indicate whether the translations are sufficiently accurate for practical use without introducing potential risks. Future work will involve evaluating clinical risks, following the approach of Mehandru et al. (2023).

References

Marc A. Adams and Terry L. Conway. 2014. *Eta Squared*, pages 1965–1966. Springer Netherlands, Dordrecht.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Seth Aycock and Rachel Bawden. 2024. Topic-guided example selection for domain adaptation in LLM-based machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 175–195, St. Julian's, Malta. Association for Computational Linguistics.

Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. 2021. A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 135–142, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, page 1877–1901. Curran Associates, Inc.

Brittany M. C. Chan, Jeanine Suurmond, Julia C. M. Van Weert, and Barbara C. Schouten. 2024. Uncovering communication strategies used in language-discordant consultations with people who are migrants: Qualitative interviews with healthcare providers. *Health Expectations*, 27(1):e13949.

Pranjal Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in LLMs for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7384–7406, Bangkok, Thailand. Association for Computational Linguistics.

M. Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137, Copenhagen, Denmark. Association for Computational Linguistics.

Johanna Gerlach, Hervé SPECHBACH, and Pierrette Bouillon. 2018. Creating an online translation platform to build target language resources for a medical phraselator. In *Proceedings of the 40th edition of Translating and the Computer Conference (TC40)*, pages 60–65. AsLing, The International Association for Advancement in Language Technology, London (UK).

Scott L Greer. 2012. The european centre for disease prevention and control: hub or hollow core ? *Journal of health politics, policy and law*, 3737(6)(1):1001—-1030.

Barry Haddow. 2015. HimL (health in my language). In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, Antalya, Turkey. European Association for Machine Translation.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Tianxiang Hu, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA. Association for Computational Linguistics.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société vaudoise des sciences naturelles (in French)*, 37 (142):547–579.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Gwénolé Lecorvé, Hugo Ayats, Benoît Fournier, Jade Mekki, Jonathan Chevelu, Delphine Battistelli, and Nicolas Béchet. 2023. Towards the automatic processing of language registers: Semi-supervisedly built corpus and classifier for french. In *Computational Linguistics and Intelligent Text Processing*, pages 480–492, Cham. Springer Nature Switzerland.

Zeyang Liu, Ke Zhou, Jiaxin Mao, and Max L. Wilson. 2021. Posscore: A simple yet effective evaluation of conversational search with part of speech labelling. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 1119–1129, New York, NY, USA. Association for Computing Machinery.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

Long Phan, Tai Dang, Hieu Tran, Trieu H. Trinh, Vy Phan, Lam D. Chau, and Minh-Thang Luong. 2023. Enriching biomedical knowledge for low-resource language through large-scale translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3131–3142, Dubrovnik, Croatia. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Amanda Ross and Victor L. Willson. 2017. *Factorial Anova*, pages 25–29. SensePublishers, Rotterdam.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 173–185, Macau SAR, China. Asia-Pacific Association for Machine Translation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht,

Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Marco Zappatore and Gilda Ruggieri. 2024. Adopting machine translation in the healthcare sector: A methodological multi-criteria review. *Computer Speech & Language*, 84:101582.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Shaolin Zhu, Menglong Cui, and Deyi Xiong. 2024a. Towards robust in-context learning for machine translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16619–16629, Torino, Italia. ELRA and ICCL.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Yuanyi Zhu, Maria Liakata, and Giovanni Montana. 2024c. A multi-task transformer model for fine-grained labelling of chest X-ray reports. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 862–875, Torino, Italia. ELRA and ICCL.

A Example Translation Corpora

Table 2 illustrates a typical sentence from our medical dialogue dataset in French, English, and Spanish. As seen in the example, the translations are target-oriented and adapted to the communicative context: the English version uses an idiomatic rendering ("ringing noise in your ears"), while the Spanish version employs an equivalent ("zumbidos") rather than a literal calque of the French source term. This reflects the dataset's design guidelines, which emphasized the audience awareness and freedom of reformulation.

Language	Sentence
French	Pendant combien de jours avez-vous pris des médicaments contre les acouphènes ?
English	How many days did you take medicine to help with the ringing noise in your ears for?
Spanish	¿Durante cuántos días ha estado tomando medicamentos contra los zumbidos?

Table 2: Example of a medical dialogue sentence in three languages.

B Settings

B.1 Model Prompt and Design

```
Can you translate from English to French ?

Return the result in JSON format with the following schema:

{{
  "translation": {{
    "type": "string"
  }}
}}

Generate the translation for the text that appears after <<<>>>.
Do not provide explanations or additional comments. You can return only one variation.

###
Here are some examples:

English: I will give you a prescription for cortisone
French JSON: {'translation': 'je vais vous prescrire de la cortisone'}

English: I will give you a prescription for medicine with cortisone in it
French JSON: {'translation': 'je vais vous prescrire des médicaments à base de cortisone'}

English: I will give you a prescription for steroids
French JSON: {'translation': 'je vais vous prescrire des stéroïdes'}

English: I will give you a prescription for a cream
French JSON: {'translation': 'je vais vous prescrire une crème'}

###
<<<
English: I will give you a prescription for a cortisone cream. Cortisone is a steroid that helps stop swelling.
>>>
```

Figure 7: Prompt structure for in-context learning, illustrated for English-to-French text translation. The prompt provides an instruction with output schema, a few example input-output pairs in JSON format, and then the test input demarcated by special tokens.

We use Mistral-7B-Instruct v0.3, a 7-billion parameter decoder-only LLM, as the backbone. This model was chosen because at the time of experimentation it was one of the stronger openly-available instruction-tuned models. Notably, Mistral-Instruct is predominantly trained on English and lacks dedicated support for many of our languages (e.g. Tigrinya), making it a good stress-test for ICL. We access the model via HuggingFace Transformers, running in half-precision (fp16) with FlashAttention optimization for efficiency. Generation is done greedily (no sampling) to ensure deterministic outputs for a given prompt.

We construct a prompt template that includes an instruction section, a few example translation pairs, and then the input to translate. The instruction defines the task (e.g., "Translate from language X to language Y and output in a JSON format"). We enforce a JSON output schema to ensure the model's output is structured correctly. An example prompt (for English-to-French translation) is shown in Figure 7. The prompt begins with a task description and output schema specification (the schema indicates that the output should be a JSON with a "translation" field containing a string). It then says: "Here are some examples:" followed by N example pairs. Each example is formatted as:

[source language]: [source text example]

[target language] JSON: "translation": "[target text example]"

After listing the N examples, the prompt has a separator and then the actual input to be translated, marked clearly (e.g., by <<< >>>). The model is expected to produce only the JSON translation for the input without additional commentary. We found that including the language names (as in the figure) helps the model produce the output in the correct language, especially since the model is multi-lingual only through prompting. This prompt format was kept consistent in all experiments to focus on the content of examples rather than prompt wording.

C Effect of Factors on Translation Quality

Figures 9, 8 and 10 show the detailed results for the effect of the number of examples and the selection method across test sets by langauge. Each subfigure presents ChrF scores for examples drawn from different registers (BM25, LASER, RANDOM). Statistical significance between registers is indicated in the plots. These results complement the main findings in Section 4.1, providing per-dataset and per-method breakdowns that were summarised in the main text.

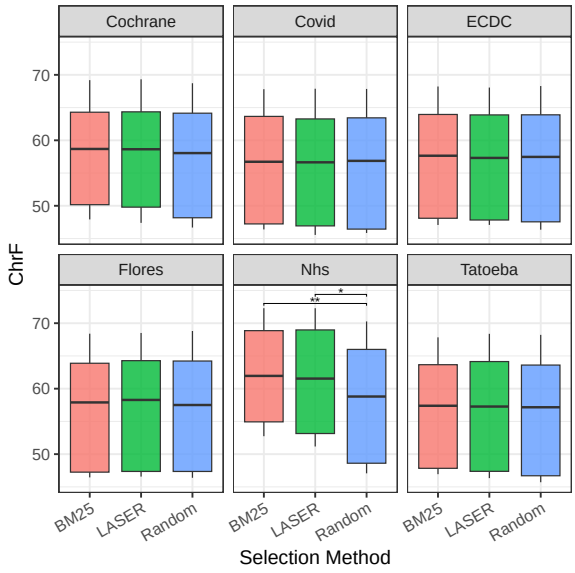


Figure 8: Effect of selection method on translation quality across different test datasets tested in NHS24. ChrF and COMET scores are shown for in-context examples drawn from different selection methods (BM25, LASER, RANDOM) on the different domain examples. Statistical significance among results is indicated by *** when $p < 0.001$, ** when $p < 0.01$ and * when $p < 0.05$.

D Linguistic Evaluation

Figures 11 and 12 present the full scatterplots for the relationship between prompt-level similarity metrics and average translation quality. Each figure corresponds to one similarity feature:

- Figure 11: **POS Distance** (cosine distance between part-of-speech distributions).
- Figure 12: **Semantic Alignment** (cosine similarity between sentence embeddings).

Spearman's ρ is shown for each plot, indicating both the strength and direction of the correlation. These figures provide the complete visual evidence underlying the correlation values reported in Section 5.

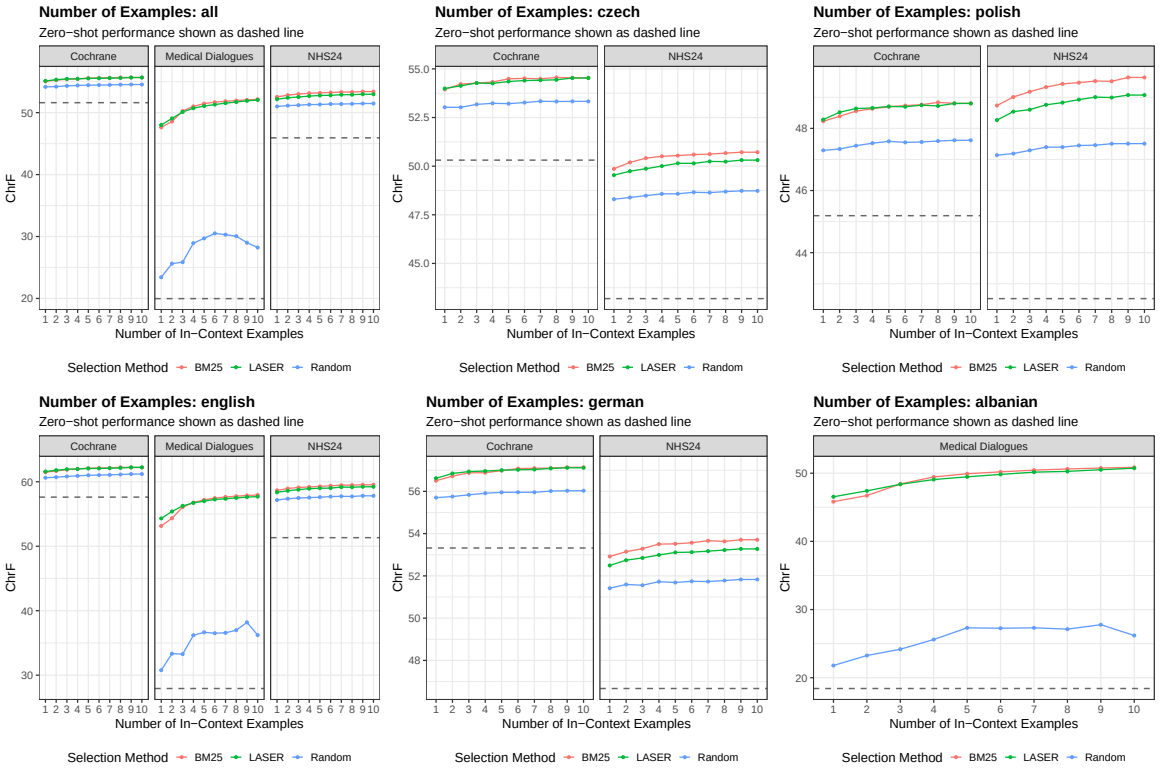


Figure 9: Effect of number of examples and selection method per test data. ChrF scores are shown for in-context examples drawn from different registers using three selection methods (BM25, LASER, RANDOM). Statistical significance between registers is indicated.

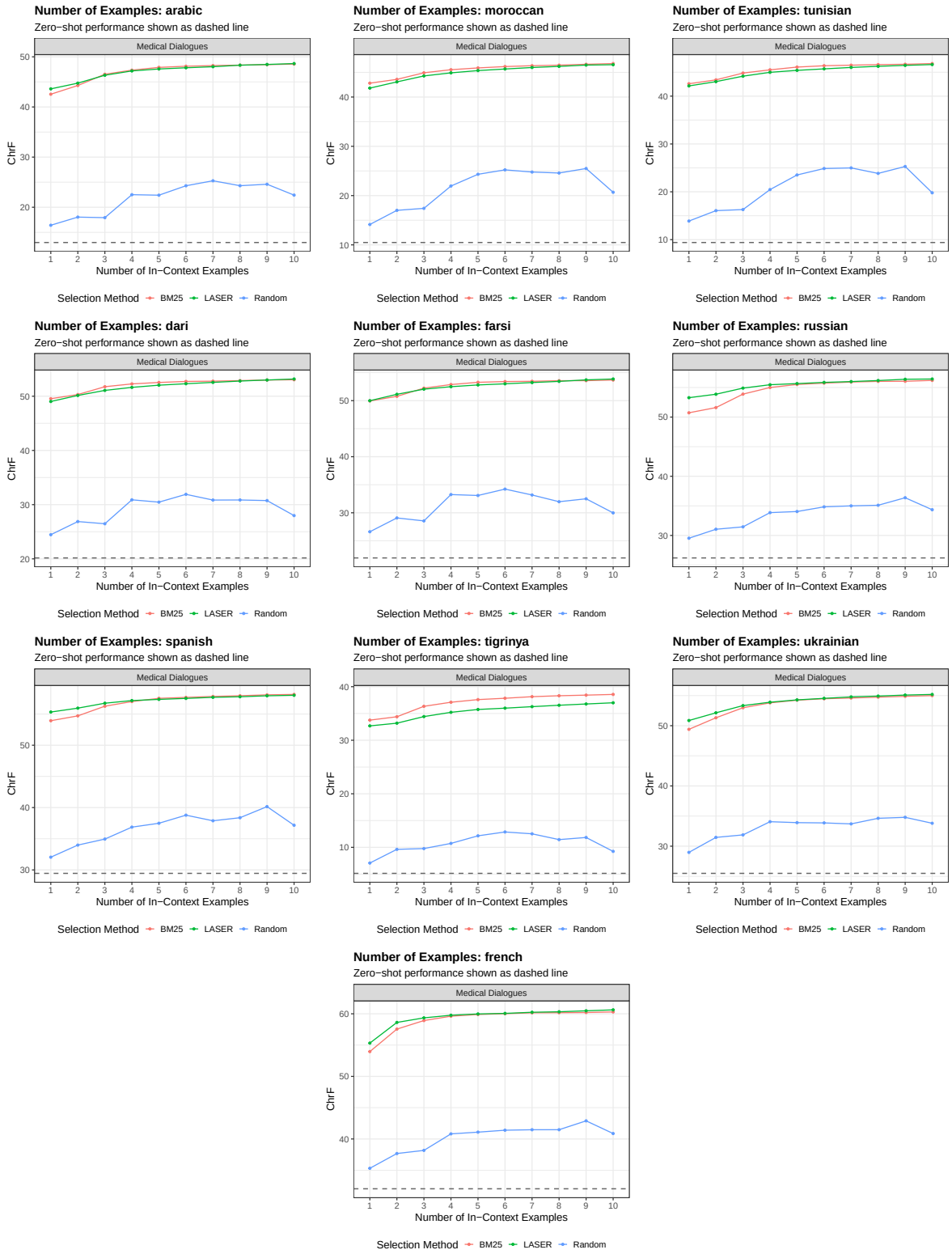


Figure 10: Effect of number of examples and selection method per test data. ChrF scores are shown for in-context examples drawn from different registers using three selection methods (BM25, LASER, RANDOM). Statistical significance between registers is indicated.

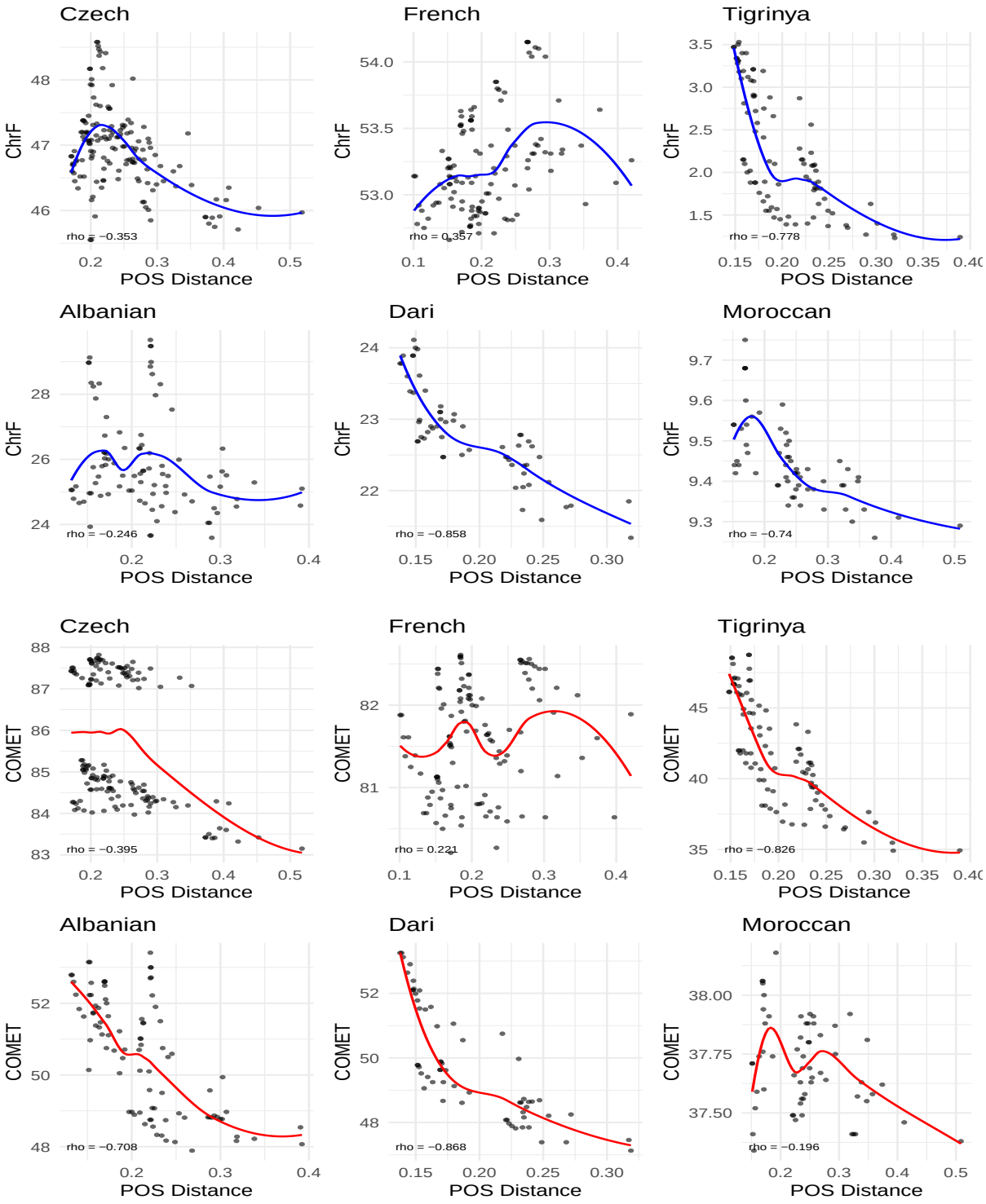


Figure 11: Scatterplots showing the relationship between prompt-level similarity metrics and average translation quality. Each plot corresponds to one of the following prompt-level measurements **POS Distance** (cosine distance between part-of-speech distributions). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

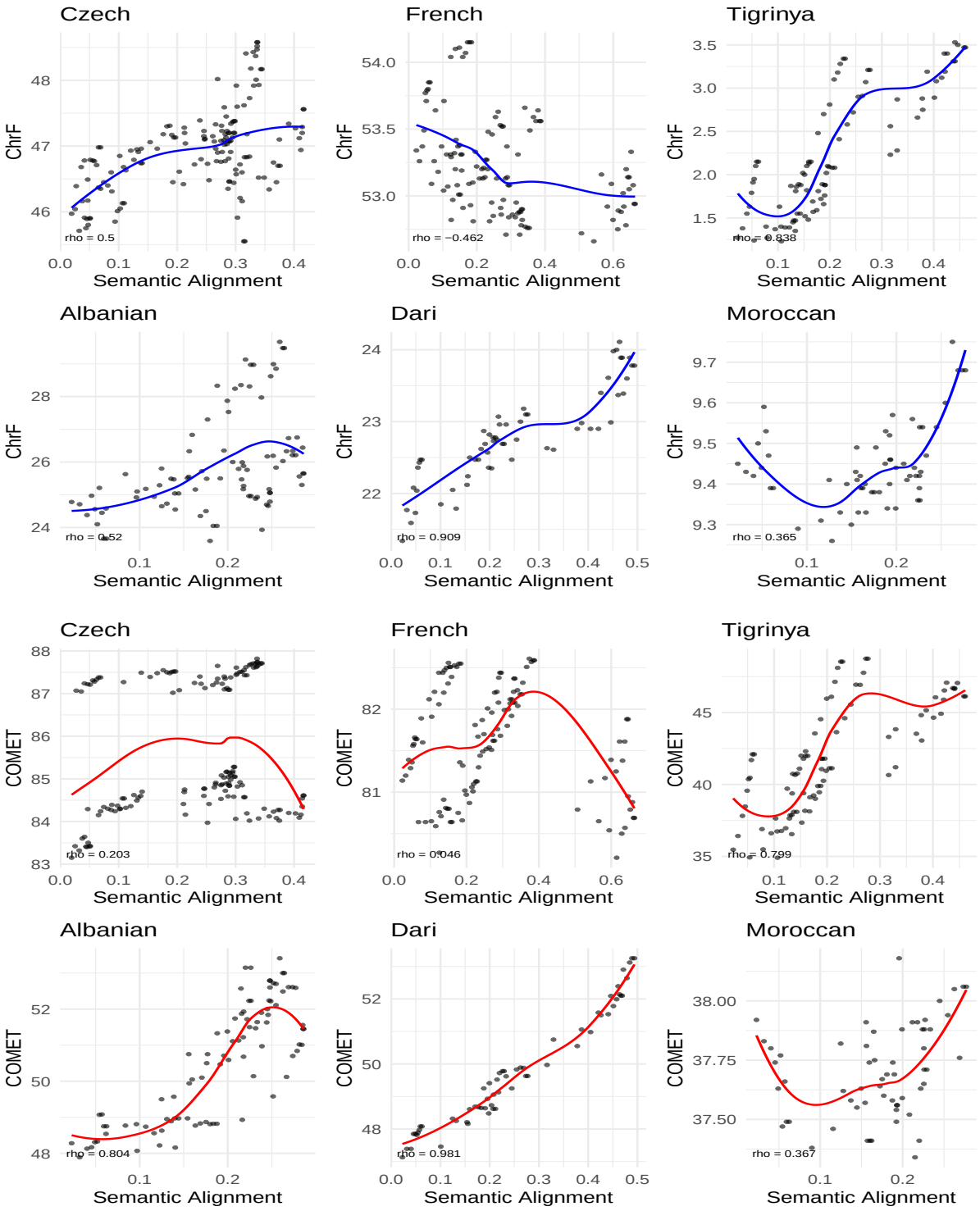


Figure 12: Scatterplots showing the relationship between prompt-level similarity metrics and average translation quality. Each plot corresponds to one of the following prompt-level measurements for **Semantic Alignment** (cosine similarity between sentence embeddings). Spearman's ρ is shown for each metric, indicating the strength and direction of correlation.

Character-Aware English-to-Japanese Translation of Fictional Dialogue Using Speaker Embeddings and Back-Translation

Ayuna Nagato **Takuya Matsuzaki**
Tokyo University of Science
1425527@ed.tus.ac.jp matuzaki@rs.tus.ac.jp

Abstract

In Japanese, the form of utterances often reflect speaker-specific character traits, such as gender and personality, through the choise of linguistic elements including personal pronouns and sentence-final particles. However, such elements are not always available in English and a character's traits are often not directly expressed in English utterances, which can lead to character-inconsistent translations of English novels into Japanese. To address this, we propose a character-aware translation framework that incorporates speaker embeddings. We first train a speaker embedding model by masking the expressions in Japanese utterances that manifest the speaker's traits and learning to predict them. The resulting embeddings are then injected into a machine translation model. Experimental results show that our proposed method outperforms conventional fine-tuning in preserving speaker-specific character traits in translations.

1 Introduction

Neural machine translation (NMT) has made remarkable progress in recent years. However, translating fictional narratives, such as novels, still poses substantial challenges. Prior work has pointed out difficulties such as preserving long-range coherence, maintaining consistent tone across chapters, and handling figurative or culturally specific expressions (Thai et al., 2022; Liu et al., 2023; Karpinska and Iyyer, 2023). One of the less-studied issues is the preservation of character-specific linguistic style, especially in the translation of dialogue.

This problem becomes particularly apparent when translating from English into Japanese. Japanese dialogue often encodes rich speaker characteristics – such as gender, personality, and social status – through a variety of linguistic devices, including first- and second-person pronouns, honorifics, and sentence-final particles. In contrast, English dialogue tends to be less explicit in expressing such traits. As a result, standard translation models often produce Japanese utterances that contradict the original speaker's identity, leading to unnatural or inconsistent character portrayals.

One major obstacle in addressing this issue is the scarcity of high-quality bilingual dialogue corpora in the literary domain. To overcome this, we employ a back-translation (Sennrich et al., 2016) strategy: we first translate a large collection of Japanese novels into English using a high-performing neural MT system. This enables us to construct a large-scale pseudo-parallel corpus of Japanese-English fictional dialogue, which serves as the foundation for training our models.

While fine-tuning translation models on such in-domain dialogue can partially alleviate the problem, it is insufficient to capture the nuanced stylistic variation required for faithful character portrayal. We hypothesize that explicitly modeling the speaker's identity can help resolve this issue.

In this work, we propose a character-aware translation model that integrates speaker embeddings into the translation process. These embeddings are learned from Japanese utterances by masking the expressions that manifest the speaker's traits and training the model to predict them, capturing latent speaker traits in a data-driven way. We inject these embeddings into a Transformer-based translation model and fine-tune it on bilingual literary dialogue data. Experimental results show that our approach produces translations that better preserve speaker-specific character traits compared to conventional fine-tuned baselines.

The contributions of this paper are as follows:

- We introduce a novel speaker embedding model tailored to Japanese dialogue.
- We incorporate these embeddings into a neural translation model.
- We utilize back-translated Japanese-English

Proceedings of the Tenth Conference on Machine Translation, pages 180–190
November 8-9, 2025

novel data to overcome the lack of existing bilingual corpora.

- We demonstrate both qualitative and quantitative improvements in preserving character consistency in Japanese translations.

2 Background: Character Expressiveness in Japanese Utterances

Japanese is a language rich in surface-level variation that reflects the speaker's social identity, personality, and emotional stance. In spoken language, particularly in literary dialogue, this variation is often encoded through the choice of **personal pronouns**, **honorific expressions**, and **sentence-final particles**. These elements do not simply convey information but actively construct the speaker's character. In this section, we outline each of these linguistic mechanisms and explain how they contribute to speaker characterization in Japanese.

2.1 Personal Pronouns

Unlike English, Japanese personal pronouns vary widely depending on the speaker's gender, formality, and social distance. Even within first- and second-person references, different pronouns evoke distinct speaker personas.

For example, for the first-person pronoun "I", speakers may choose from:

- 私 **(watashi)** : neutral or formal
- 僕 **(boku)** : typically used by polite males
- 俺 **(ore)** : rough or masculine tone
- あたし **(atashi)** : casual and feminine

and many more.

For second-person references:

- あなた **(anata)** : formal or neutral
- おまえ **(omae)** : rough, informal, sometimes aggressive
- あんた **(anta)** : casual, often used by women
- きみ **(kimi)** : gentle, sometimes condescending depending on context

2.2 Sentence-Final Particles

Sentence-final particles such as よ **(yo)**, ね **(ne)**, の **(no)**, ぞ **(zo)**, and わ **(wa)** play a key role in expressing pragmatic and emotional nuance.

- よ **(yo)** : adds emphasis or confidence
- ね **(ne)** : invites agreement or shared understanding
- の **(no)** : softens a statement, often used by female speakers
- ぞ **(zo)**, ぜ **(ze)** : express strong masculine emphasis
- わ **(wa)** : indicates a feminine or classical tone depending on usage

2.3 Honorifics and Politeness Levels

Japanese exhibits a highly stratified system of honorifics, including respectful, humble, and polite forms. These levels express not only social hierarchy but also character traits in literary dialogue.

Politeness can be expressed by an auxiliary verb or a light verb:

- です **(desu)**/ます **(masu)** : basic politeness
- ございます **(gozaimasu)** : highly respectful
- くださる **(kudasaru)** : humble expressions

as well as the choice of a verb:

- 食べる **(taberu)** ↔召し上がる **(mesiagaru)** : normal ↔polite form of "eat"
- 言う **(iu)** ↔おっしゃる **(ossyaru)** : normal ↔polite form of "say"

2.4 Variation in Utterances Reflecting Speaker Character

The above linguistic features often appear together, shaping the overall tone and personality of the speaker. Table 1 shows different utterances of "Who are you?" that reflect various speaker identities through the use of pronouns, honorifics, and sentence-final particles.

Although all the examples convey the same core meaning, the speaker's personality, social stance, and emotional intensity vary drastically. In English–Japanese translation of literary dialogue, these nuances have to be properly differentiated by the

Table 1: Examples of speaker-dependent utterance variation in Japanese

Utterance	Pronoun	Honorifics	Final Particle	Character Impression
あなたはどなた**ですか**？	あなた (anata)	です (desu)	か (ka)	Formal, respectful
おまえ、誰だ？	おまえ (omae)	none	none	Rough, masculine
あんた、誰**よ**？	あんた (anta)	none	よ (yo)	Strong-willed female
きみは誰な**の**？	きみ (kimi)	none	の (no)	Friendly, gentle

choices of linguistic features according to the character, even if the source English expression is the same. Conventional systems often produce translations that fail to align with the character's original persona. This motivates our approach to incorporate speaker embeddings into translation to better preserve character-specific traits.

3 Method

Figure 1 illustrates the architecture used for constructing speaker embeddings and utilizing them in English-Japanese translation. We detail the method in what follows.

3.1 Speaker Embedding Construction

To incorporate speaker-specific characteristics into the translation process, we construct a speaker embedding model trained on Japanese literary dialogue. The aim is to learn embeddings that capture the personality traits expressed in each character's speech. The training proceeds as follows:

Step 1: Creating a Japanese-English Parallel Corpus Due to the scarcity of parallel corpora of Japanese and English novels, we create a pseudo-parallel corpus by translating 13,772 Japanese novels from Aozora Bunko[1] into English. We translated these novels by using a Transformer-based large model trained on JParaCrawl v3 (Morishita et al., 2022). This pseudo-parallel corpus is also used in the training of the speaker-aware translation model described in Section 3.2.2.

Step 2: Speaker Identification Using Stanford CoreNLP (Manning et al., 2014), we extract speaker-utterance pairs and the sentences where the subject is one of the speakers from the English translations. Specifically, we extract (i) utterances and their speakers through quote attribution, and (ii) declarative sentences where one of the speakers in a novel is marked as the subject (nsubj). These English sentences are then aligned with their Japanese counterparts to create bilingual dialogue pairs.

Step 3: Speaker-Sensitive Masking We mask parts of each Japanese utterance that tend to encode speaker-specific traits using the following rules:

- **Pronouns (1st and 2nd person):** Tokens tagged as "pronoun" by MeCab morphological analyzer (Kudo et al., 2004) are replaced with [MASK].
- **Sentence-final particles:** Tokens tagged as "sentence-final particle" are masked.
- **Honorific expressions:** Polite or honorific expressions, including verbs and sentence-final copulas such as です (desu) and ます (masu) are masked. If none of these forms are present at the sentence end, an empty [MASK] token is inserted to maintain output consistency.

Step 4: Embedding Extraction To obtain the embedding of speaker X, all utterances by X and the sentences with X as the subject are extracted from an English novel, concatenated using [SEP] as separators, truncated to 512 tokens, and input to English BERT [2]. The resulting [CLS] token embedding is used as the speaker's embedding vector.

Step 5: Training the Speaker Embedding Model Masked Japanese utterances are input to Japanese BERT [3]. Each output token vector is combined with the speaker's English BERT embedding by vector addition and passed through the language modeling head to predict the masked tokens. Cross-entropy loss is computed only at masked positions. Figure 2 presents an overview of Step 4 and 5. All components—Japanese BERT, English BERT, and the language modeling head—are jointly fine-tuned.

For consistency with the downstream translation model, we also fine-tune the speaker embedding model using the aligned NICT corpus described in Section 3.2.2.

[1] https://www.aozora.gr.jp

[2] https://huggingface.co/google-bert/bert-base-uncased

[3] https://huggingface.co/tohoku-nlp/bert-base-japanese-v3

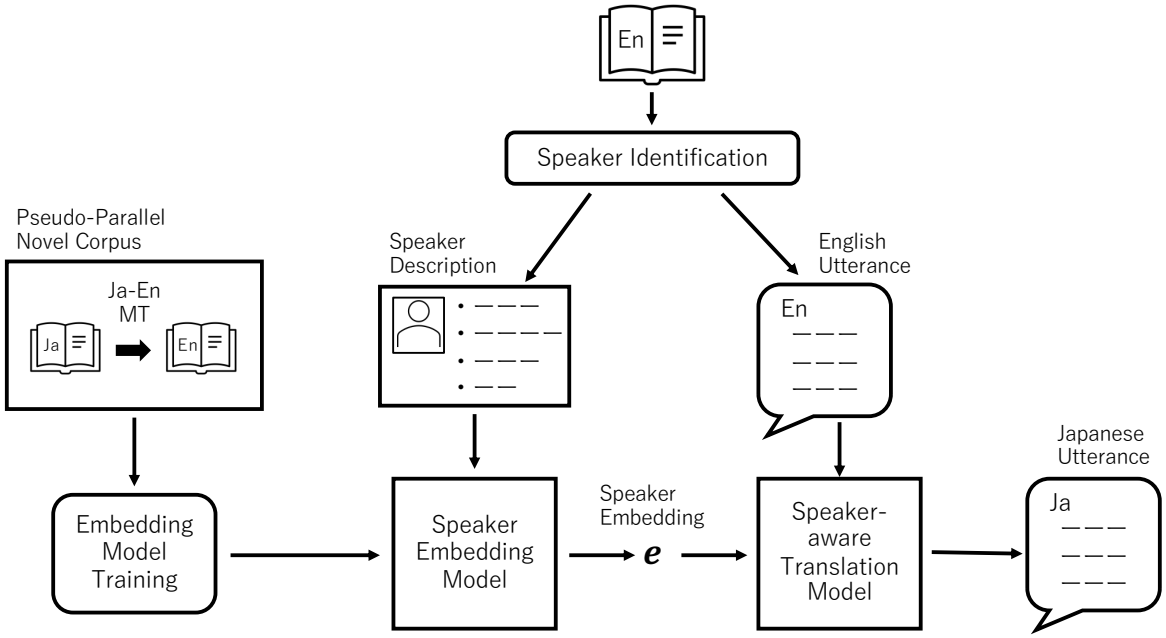


Figure 1: Overview of our proposed method.

For optimization, we used AdamW with hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$, and learning rate of 10^{-5}. The batch size was set to 32. We chose the model achieved minimum loss on the validation set within five epochs.

3.2 Speaker-Aware Translation Model

3.2.1 Model Architecture

Our translation model adopts a Transformer encoder-decoder architecture. To incorporate character-specific style into the output, we inject speaker embeddings into the decoder. Specifically, the speaker embedding is added to the hidden states of the decoder's last layer before the final linear projection:

$$z_i' = z_i + e_s$$

where z_i is the decoder hidden state at time step i, and e_s is the speaker embedding vector. This additive integration enables the model to condition generation on speaker traits such as personality or social role.

3.2.2 Training Procedure

We employ a three-stage training process designed to balance general translation quality with speaker-sensitive stylistic control. These stages are: (1) pretraining of a general-domain English-to-Japanese model, (2) training on back-translated literary dialogue, and (3) fine-tuning on sentence-aligned, human-translated dialogue from novels.

Pretraining on JParaCrawl (English to Japanese) We initialize our model using a Transformer-based large model trained on JParaCrawl v3 (Morishita et al., 2022), in the English-to-Japanese direction. This model provides a strong general translation foundation but does not incorporate speaker-specific information. It serves as the backbone for subsequent adaptation.

Training with Back-Translated Literary Dialogue As described in Section 3.1, we utilize a pseudo-parallel corpus of Japanese novels and their translations to English. We use this back-translated dataset—composed of Japanese original dialogue and its English translation—to fine-tune our English-to-Japanese translation model with speaker embeddings. The Japanese side contains rich stylistic expressions, and the English side provides automatically extracted speaker labels via quote attribution and syntactic parsing.

This stage allows the model to learn how speaker-specific stylistic features in Japanese correspond to the more neutral English dialogue, and how to generate speaker-aware Japanese output based on speaker embeddings.

For optimization, we used Adam with hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$, and learning rate of 5×10^{-5}. The batch size was set to 6000 tokens. We trained up to 20k updates and averaged the last eight checkpoints for the final model.

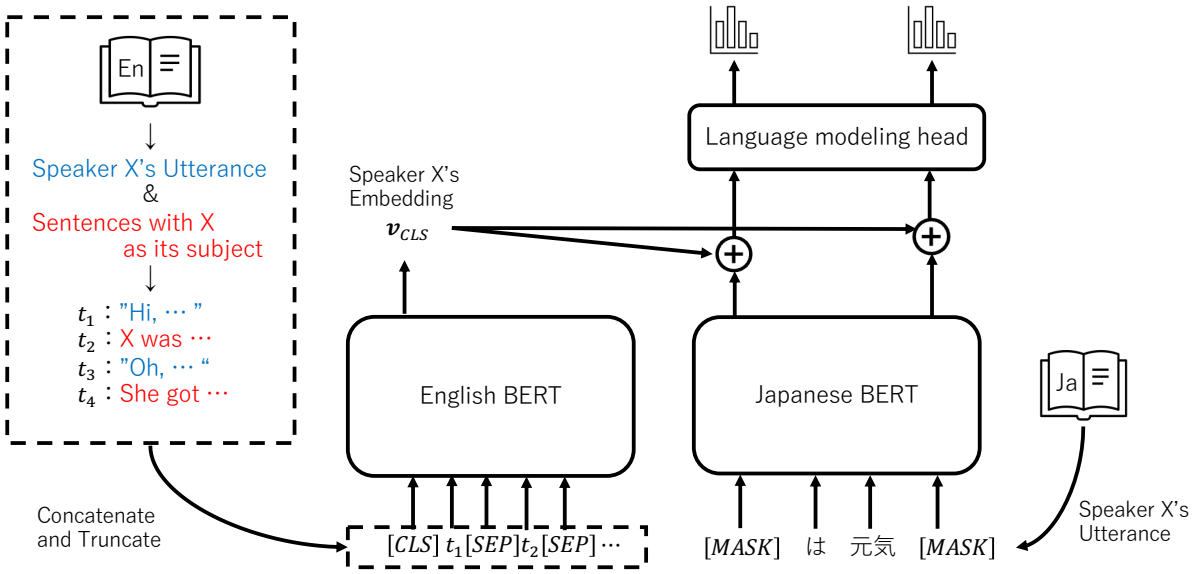


Figure 2: Architecture of the speaker embedding model. The masked Japanese input is processed by Japanese BERT, while the speaker embedding from English BERT is integrated during token prediction.

Fine-Tuning with Sentence-Aligned Modern Dialogue Although the back-translated Aozora data is rich in stylistic variation, it tends to reflect older literary styles. To adapt the model to contemporary Japanese, we fine-tune it using a small set of manually sentence-aligned English-Japanese novel data that was developed by Utiyama and Takahashi (2003) and is distributed by the National Institute of Information and Communication Technology (NICT). We henceforth call this data the NICT corpus. It mostly consists of modern fictional texts aligned at the sentence level, but without explicit speaker annotations.

We extract speaker information automatically using Stanford CoreNLP on the English side, as described in Section 3.1, and generate speaker embeddings using our trained embedding model. This final step improves fluency, modernity, and alignment with contemporary character dialogue styles.

For optimization, we used Adam with hyperparameters of $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-8}$, and a smaller learning rate of 10^{-5} to avoid overfitting on the limited NICT corpus. The batch size was set to 2000 tokens. We trained up to 2k updates and averaged the last three checkpoints for the final model.

Data Splits All training, validation, and test sets for the English-to-Japanese translation consist exclusively of utterances from dialogue segments, with narrative text excluded (except as the input to the speaker embedding model). For both the back-translated Aozora corpus and the aligned NICT corpus, we split the data into training, validation, and test sets in an 8:1:1 ratio. Importantly, the split is done on a per-work basis (i.e., by novel title) to avoid information leakage across sets. Furthermore, to ensure corpus independence, we exclude from the Aozora corpus any works that are also included in the NICT corpus.

4 Experiments

This section presents a comprehensive evaluation of the proposed speaker-aware translation model. We first conduct qualitative and quantitative manual analyses to assess how incorporating speaker embeddings affects character-sensitive translation aspects such as pronoun choice and sentence-final particles. Next, a case study on a single character from a literary work examines consistency in character voice over multiple utterances. Additionally, we perform automatic evaluation using BLEU, ChrF, and COMET-22 scores on dialogue-heavy test data to quantify improvements over baseline models. Finally, we analyze the learned speaker embeddings via Principal Component Analysis (PCA) to interpret the linguistic and stylistic features captured in the embedding space.

Together, these evaluations demonstrate the effectiveness of integrating speaker embeddings in enhancing the fidelity and expressiveness of literary dialogue translation.

4.1 Quantitative Analysis: Manual Evaluation of Speaker-Sensitive Translation

To qualitatively and quantitatively evaluate the impact of incorporating speaker embeddings, we conducted a manual analysis of sampled outputs. From the 1,436 utterances in the NICT test set, we randomly selected 150 utterances and compared translations produced by two systems:

- **Baseline**: A standard Transformer model pretrained on JParaCrawl v3 and fine-tuned on back-translated Aozora data followed by fine-tuning on the NICT corpus.
- **Proposed**: The same model architecture and training procedure, but augmented with speaker embeddings during training and inference.

We evaluated each output along five criteria:

- **Pronouns**: Whether first- and second-person pronouns exactly matched the reference translation (surface differences such as 私 (watashi) vs わたし (watashi) were treated as equivalent).
- **Sentence-final particles**: Whether sentence-final particles exactly matched the reference translation.
- **Honorific usage**: Whether honorific or polite expressions were used when the reference translation employed them.
- **Intra-utterance consistency**: Whether a single utterance maintained consistent character traits (e.g., not mixing 僕 (boku) and 私 (watashi)).
- **Translation errors**: Whether the output contained critical errors such as omissions, repetitions, or untranslated segments.

The results are summarized in Table 2. For pronouns, sentence-final particles, and honorifics, we report accuracy (percentage of exact matches with the reference). For consistency and translation errors, we report the number of problematic utterances out of the 150 sampled.

Manual evaluation results, as shown in Table 2, demonstrate several notable improvements brought by our proposed speaker-aware translation model.

First, the accuracy of pronoun translation significantly increased from 25.6% in the baseline to 61.6% in our model. This suggests that incorporating speaker embeddings greatly improves the model's ability to select appropriate first- and second-person pronouns, which are crucial in reflecting character-specific traits. Similarly, accuracy on sentence-final particles improved from 40.5% to 52.5%, despite the strict criterion of requiring exact matches. In fact, our manual inspection revealed that some minor mismatches—such as わ (wa) vs. よ (yo) or だ (da) vs. ぞ (zo) —still produce utterances with similar character impressions, implying that the model's improvement may be underrepresented by exact matching metrics alone.

Interestingly, performance on honorific forms remained unchanged (28.1% accuracy in both models). One possible explanation is that honorific expression tends to depend more on the social relationship between the speaker and the hearer, rather than on the speaker's character identity alone. This highlights a potential limitation of our speaker-only embedding approach in capturing such pragmatic nuances.

The number of consistency errors —-such as inconsistent use of personal pronouns within the same utterance—- decreased from seven to zero. This indicates that the proposed model contributes to maintaining character consistency at the utterance level. Furthermore, the number of critical translation errors, including untranslated segments and repeated phrases, was reduced from four to zero.

Taken together, these findings support the effectiveness of integrating speaker embeddings in improving character-sensitive aspects of translation, especially for pronoun and sentence-final particle choices, while also enhancing consistency of character traits.

4.2 Case Study: Fatty Coon in *The Tale of Fatty Coon*

To further analyze the effect of our method on maintaining the consistency of character's traits, we conducted a case study on Fatty Coon, the protagonist of *The Tale of Fatty Coon* by Arthur Scott Bailey. This character is portrayed as an energetic young boy raccoon, often using casual language such as ぼく (boku) and sentence-final particles like よ (yo) in the Japanese translation.

Out of 74 utterances attributed to Fatty Coon (based on CoreNLP speaker tagging), 9 were misattributed. We excluded these and manually analyzed

Table 2: Manual evaluation results on 150 randomly sampled utterances from the NICT test set. Accuracy is reported for categorical items (pronouns, sentence-final particles, honorifics), and raw counts are reported for consistency errors and critical translation errors.

Model	Pronouns	Final Particles	Honorifics	Consistency Errors	Translation Errors
Baseline	25.6% (32/125)	40.5% (98/242)	28.1% (9/32)	7	4
Proposed	**61.6% (77/125)**	**52.5% (127/242)**	28.1% (9/32)	**0**	**0**

the remaining 65 utterances.

We found that the baseline model frequently produced outputs that were inconsistent with the character's personality. Specifically, 29 out of the 65 utterances (44.6%) included language that was too formal or feminine, such as the use of 私 (watashi) or sentence-final particles like わ (wa), or even polite verb forms. In contrast, our proposed method produced consistent outputs aligned with the character's casual and boyish tone in nearly all cases, with only 9 utterances showing mismatches (e.g., use of honorifics).

Table 3 shows representative examples. In each case, the proposed method produces translations that are closer to the reference translation and consistent with the intended persona of Fatty Coon.

4.3 Automatic Evaluation on NICT Dialogue Segments

To complement our manual evaluation, we conducted automatic evaluation using BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and COMET-22 (Rei et al., 2022) on the NICT test data, comprising 1,436 Japanese utterances. We report BLEU scores computed using `sacreBLEU` with `--tokenize=intl` option. COMET is a neural-based metric trained to predict human direct assessment scores, and has been shown to correlate more strongly with human evaluation than surface-overlap metrics such as BLEU (Rei et al., 2020).

Table 4 provides the results. The absolute BLEU and ChrF scores appear low, especially compared to typical scores reported in general domain English-to-Japanese translation. However, our task differs significantly in both content and style: the data is literary dialogue, which contains diverse speaker-specific expressions, idiosyncratic phrasing, and multiple valid translations. In such settings, surface-form overlap metrics like BLEU often underestimate translation quality (Toral and Way, 2018; Mathur et al., 2020; Thai et al., 2022).

Despite the limitations of these metrics in capturing speaker-specific style or consistency, the proposed method outperforms both the base model and the fine-tuned baseline by a noticeable margin on both BLEU and ChrF. This suggests that introducing speaker-aware information helps produce translations that better align with the reference utterances even in automatic evaluation metrics. In the case of BLEU score, the larger gain compared to the fine-tuned baseline also supports our hypothesis that speaker traits contribute to reducing ambiguities in character-driven translation.

In addition, the COMET-22 scores also show that our proposed method achieves the highest performance among the compared systems (0.802 vs. 0.778 for the baseline). This further supports that incorporating speaker information not only improves surface-level similarity but also yields translations that are semantically closer to human references, in line with human judgments of adequacy.

To assess the reliability of these improvements, we conducted paired bootstrap resampling tests. BLEU differences between the baseline and our proposed model were not statistically significant ($p = 0.14$). However, both ChrF ($p < 0.01$) and COMET ($p < 0.05$) confirmed that the improvements of the proposed model over the baseline are statistically significant. These results suggest that while BLEU may underestimate gains in this task, stronger metrics such as ChrF and COMET provide more robust evidence that speaker-aware information improves translation quality.

Nonetheless, given the nature of the task, we emphasize that manual evaluation and qualitative analysis (as described in the previous subsections) provide a more reliable assessment of character expressiveness and speaker consistency in translation.

4.4 Speaker Embedding Visualization via PCA

To investigate the structure of the learned speaker embeddings, we apply Principal Component Analysis (PCA) to the speaker vectors extracted from the test portion of the back-translated Aozora corpus. This test set was held out during model training.

We first applied our trained speaker embedding model to the machine-translated English versions

Table 3: Example translations of utterances by Fatty Coon. The proposed method produces outputs more consistent with the character's persona.

Source	Reference	Baseline	Proposed
"I'd like to eat all the corn in the world."	世界中のとうもろこしを全部食べちゃいたい**よ (yo)**。	世界中のとうもろこしをみんな食べてみたい**わ (wa)**。	世界中のとうもろこしがみんな食べたい**よ (yo)**。
"Look, Mother!"	見て、お母さん！	**ご ら ん な さ い (Honorifics)**、お母さん！	見てよ、おかあさん！
"Maybe you don't think I heard him screech–"	**ぼく (boku)** があいつの絶叫を聞いたっていうのも―――	たぶん、**わたし (watashi)** が金切り声を聞いたんじゃないと思うだろう――	**僕 (boku)** が金切声を聞いたと思わないだろう――

Table 4: Automatic evaluation scores on NICT test data

Model	BLEU	ChrF	COMET-22
JParaCrawl (pre-FT)	3.8	14.8	0.776
Baseline	3.4	16.7	0.778
Proposed	**5.7**	**18.2**	**0.802**

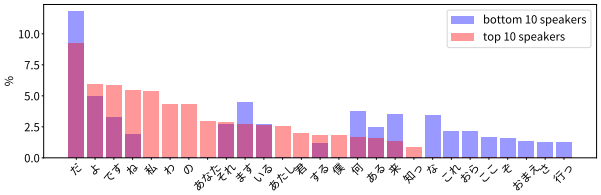


Figure 3: Proportion of [MASK] tokens in utterances of speakers with top and bottom PC Scores

of the test set novels. Each character's English utterances were concatenated and processed using BERT to produce a fixed-size speaker embedding, as described in Section 3.1. As a result, we obtained embeddings of 6,449 speakers.

We then performed PCA on these embeddings and analyzed the first principal component. Table 5 lists the top and bottom 10 characters based on their scores on the first principal component, along with their associated works and authors. We found that characters with high component scores tend to be female, while those with low scores tend to be male.

This interpretation is supported by examining the masked tokens in their utterances (Figure 3). Tokens like わ (wa), の (no), あなた (anata), and あたし (atashi) appear frequently in the utterances of characters with high PCA scores—expressions that are stereotypically feminine in Japanese. In contrast, utterances from characters with low PCA scores more often contain ぞ (zo), おまえ (omae), and おら (ora), which are typically masculine expressions.

Moreover, we observe that characters with similar component scores often come from the same literary work or author. This pattern suggests that the speaker embeddings encode not only individual character traits but also stylistic patterns associated with particular authors.

These results indicate that our speaker embedding space captures interpretable dimensions related to gendered language use and authorial style in literary dialogue.

5 Discussion

The proposed speaker-aware translation model showed improvements in preserving speaker-specific linguistic traits. However, several limitations and potential directions for future work have emerged.

Accuracy of Speaker Attribution. In this study, we relied on speaker attribution results from Stanford CoreNLP. As observed in the case study of "Fatty Coon", speaker misattribution occurred in 9 out of 74 utterances. These errors can directly affect the quality of the speaker embeddings, potentially leading to unnatural or inconsistent translations.

Effect of Character Description Length on Embedding Quality. To construct speaker embeddings, we used English character descriptions—specifically, sentences where the character was the subject and the character's utterances—from the beginning of each work. These sentences were concatenated until reaching the maximum input length of the English BERT model. As a result, the amount of contextual information varied depending on how much text was available for each character. We have not yet conducted a systematic analysis of how the amount or content of these character-descriptive sentences affects the quality of the embeddings or the final translation output. Investigating the impact of description length and exploring methods to prioritize more informative sentences (e.g., those rich in personality cues) may help enhance consistency,

Table 5: Top and bottom characters on the first PCA component

Bottom (Low PC score)			Top (High PC score)		
Speaker	Gender	Work	Speaker	Gender	Work
Kasuke	Male	“Irefuda” by Kikuchi Kan	Kuroe	Female	“Charako-san” by Hisao Juran
Koshu	Male	“Dai-bosatsu toge” by Nakazato Kaizan	her	Female	“Charako-san” by Hisao Juran
Iori	Male	“Zenigata Heiji” by Nomura Kodo	Yoshie	Female	“Charako-san” by Hisao Juran
his	Male	“The Escape of Terasaka Kichiyemon” by Naoki Sanjugo	Noriko	Female	“Sugiko” by Miyamoto Yuriko
Yasuke	Male	“The Woman Who Stepped on a Shadow” by Okamoto Kido	Haruko	Female	“Nozarashi” by Toyoshima Yoshio
Isuke	Male	“Quick-Eared Sanji” by Hayashi Fubo	Suzue	Female	“A History of a Couple” by Kishida Kunio
Iori	Male	“Miyamoto Musashi” by Yoshikawa Eiji	her	Female	“This Morning’ s Snow” by Miyamoto Yuriko
Yoriharu	Male	“The Armor of Asahi” by Kunieda Shiro	Charako	Female	“Charako-san” by Hisao Juran
his	Male	“On Leisure” by Itami Mansaku	his	Male	“Sugigaki” by Miyamoto Yuriko
Yamada	Male	“My Private Taiheiki” by Yoshikawa Eiji	Madam	Female	“The Shadowless Criminal” by Sakaguchi Ango

especially for characters with limited text.

Limitations of Back-Translation Quality. One challenge we observed is that the back-translation process used to construct the training data—specifically, translating Japanese to English and then back to Japanese—sometimes produces unnatural Japanese sentences. In our pipeline, the major source of noise stems from the Japanese-to-English translation step. When this translation is inaccurate, it leads to poor-quality pseudo-parallel data, which in turn affects the quality of the final English-to-Japanese translation model. While using manually curated parallel data would be ideal, such data is extremely limited, especially for literary dialogue. Future work may improve the overall quality by employing stronger Japanese-to-English translation models, or by integrating automatic quality filtering mechanisms to reduce the impact of noisy samples.

6 Related Work

6.1 Speaker Attribution in Narrative Texts

Speaker attribution—the task of identifying who is speaking in a given utterance—is a crucial preprocessing step for speaker-aware translation. In narrative texts such as novels, explicit speaker tags are often absent, requiring automatic identification based on linguistic cues. For English texts, tools such as Stanford CoreNLP provide heuristic-based speaker tagging, but they often fail in the presence of figurative or indirect speech.

Speaker attribution in Japanese poses additional challenges due to frequent subject omission, flexible word order, and the use of sentence-final particles that vary by speaker. Ishikawa et al. (2024) addressed this by leveraging grammatical and contextual features to estimate speaker identity in Japanese novels. Zenimoto and Utsuro (2022) proposed a method for identifying the speakers of quoted utterances in Japanese novels using a gender classification model.

6.2 Machine Translation for Literary Texts

Discourse-level literary translation remains one of the most demanding tasks in natural language processing. Unlike general-domain texts, literary works require models to handle complex semantic phenomena such as figurative language, long-range dependencies, character voice, and culturally grounded expressions (Pang et al., 2025). These aspects place high demands on translation systems, which must not only be accurate but also preserve subtle stylistic and narrative consistency.

While recent progress in large language models (LLMs) has enabled strong performance on many NLP tasks, training or fine-tuning models specifically for literary translation remains costly and resource-intensive. In response to these challenges, the WMT2023 Shared Task on Discourse-Level Literary Translation was launched, highlighting the need for models that go beyond sentence-by-sentence translation (Wang et al., 2023). Results from the shared task demonstrated that even state-of-the-art systems struggle to maintain coherence, tone, and character consistency across longer texts. These findings suggest that further advances are needed in integrating discourse-level information and stylistic modeling, particularly for literature.

7 Conclusion

This paper proposed a speaker-aware machine translation framework aimed at preserving character-specific expressions in Japanese literary dialogue. By constructing speaker embeddings from English descriptions of each character and incorporating them into the translation model, our method promotes more consistent and personality-aligned outputs.

We evaluated our approach using the NICT English-Japanese translation alignment dataset. Manual analysis showed that our method improves the consistency of personal pronouns and sentence-

final particles, which are strongly associated with character identity. However, the use of honorifics did not improve as clearly, likely because honorific usage depends more on social context—such as the relationship between the speaker and hearer—than on character traits alone.

In future work, we plan to explore the integration of situational context (e.g., dialogue participants and relationships), adopt higher-quality translation models, and refine our speaker recognition pipeline to further enhance the character consistency and fluency of literary dialogue translation.

References

Kazuki Ishikawa, Kohei Ogawaa, and Satoshi Sato. 2024. Speaker identification using speech-style encoder. *Journal of Natural Language Processing*, 31(3):894–934.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Toral and Andy Way. 2018. *What Level of Quality Can Neural Machine Translation Attain on Literary Text?*, pages 263–287. Springer International Publishing, Cham.

Masao Utiyama and Mayumi Takahashi. 2003. English-japanese translation alignment data. https://www2.nict.go.jp/astrec-att/member/mutiyama/align/index.html.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Yuki Zenimoto and Takehito Utsuro. 2022. Speaker identification of quotes in Japanese novels based on gender classification model by BERT. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 126–136, Manila, Philippines. Association for Computational Linguistics.

DTW-Align: Bridging the Modality Gap in End-to-End Speech Translation with Dynamic Time Warping Alignment

Abderrahmane Issam **Yusuf Can Semerci** **Jan Scholtes** **Gerasimos Spanakis**
Department of Advanced Computing Sciences
Maastricht University
{abderrahmane.issam, y.semerci, j.scholtes, jerry.spanakis}@maastrichtuniversity.nl

Abstract

End-to-End Speech Translation (E2E-ST) is the task of translating source speech directly into target text bypassing the intermediate transcription step. The representation discrepancy between the speech and text modalities has motivated research on what is known as *bridging the modality gap*. State-of-the-art methods addressed this by aligning speech and text representations on the word or token level. Unfortunately, this requires an alignment tool that is not available for all languages. Although this issue has been addressed by aligning speech and text embeddings using nearest-neighbor similarity search, it does not lead to accurate alignments. In this work, we adapt Dynamic Time Warping (DTW) for aligning speech and text embeddings during training. Our experiments demonstrate the effectiveness of our method in bridging the modality gap in E2E-ST. Compared to previous work, our method produces more accurate alignments and achieves comparable E2E-ST results while being significantly faster. Furthermore, our method outperforms previous work in low resource settings on 5 out of 6 language directions. [1]

1 Introduction

End-to-End Speech Translation (E2E-ST) is the task of translating speech in a source language directly into text in a target language. E2E-ST gained success and attention as an alternative to cascaded solutions where an Automatic Speech recognition (ASR) and a Machine Translation (MT) models are combined (Tang et al., 2021; Ye et al., 2022; Fang et al., 2022; Ouyang et al., 2023; Zhou et al., 2023; Le et al., 2023; Zhang et al., 2024, 2025). Cascaded solutions benefit from abundant ASR and MT data but might suffer from error propagation and high latency, which can be solved by E2E-ST.

However, training E2E-ST models is not straightforward due to the representation discrepancy between the speech and text modalities. Previous work has achieved state-of-the-art results by aligning speech and text representations at the word or token level, either using an alignment tool (Ouyang et al., 2023; Fang et al., 2022) or by generating the alignment automatically during training (Zhou et al., 2023; Zhang et al., 2023b). The closest to our work is Cross-modal Mixup via Optimal Transport (CMOT), which uses optimal transport for finding speech and text alignments. Although CMOT achieves state-of-the-art results, it does not guarantee producing monotonic alignments or ensure that each text token is assigned to at least one frame. This contradicts the expected structure of speech-text alignment and can lead to noisy alignments. Furthermore, CMOT introduces a significant training time overhead.

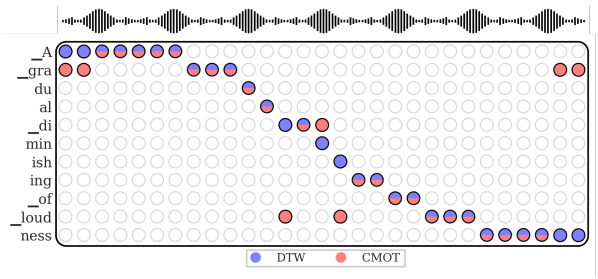


Figure 1: We show an example alignment from DTW (Ours) vs. CMOT. The figure shows that unlike CMOT, DTW guarantees generating monotonic alignments and that all tokens are aligned. In contrast, CMOT failed to align the tokens "min" and "ish" to any frames.

In this work, we introduce **DTW-Align**, a method for aligning speech and text embeddings during training using an adaptation of Dynamic Time Warping (Sakoe and Chiba, 1978). Figure 1 shows an example alignment generated using DTW-Align and CMOT, which illustrates that our method generates monotonic alignments and guarantees that all tokens are aligned, while CMOT does not. We demonstrate the effectiveness of our method

[1] https://github.com/issam9/DTW-Align

Proceedings of the Tenth Conference on Machine Translation, pages 191–199
November 8-9, 2025

in bridging the modality gap with mixup training (Fang et al., 2022; Zhou et al., 2023). Similarly to previous work (Zhou et al., 2023), we train on a mixup of aligned speech and text representations, however, instead of discretely selecting either a speech or a text embedding, we linearly interpolate speech and text embeddings (Zhang et al., 2018). Our experiments show that our method is faster and produces more accurate alignments. Furthermore, it achieves comparable results to CMOT on 6 language directions from the CoVoST2 dataset, while training significantly faster. We also evaluate our method in a low resource setting where training can be more vulnerable to alignment noise, and we show that our method leads to a statistically significant improvement over CMOT in 5 out of 6 language directions.

2 Related Works

Bridging the Modality Gap: The discrepancy between the source and target modalities (i.e. speech and text respectively) has motivated multiple works on what is termed bridging the modality gap (Liu et al., 2019; Han et al., 2021; Fang et al., 2022), where the goal is to build a shared semantic space between the speech and text modalities. Aligning speech and text either based on an alignment tool (Fang et al., 2022; Ouyang et al., 2023) or dynamically during training (Zhang et al., 2023b; Zhou et al., 2023) was shown to achieve state-of-the-art results. Our work goes in this direction, by improving the accuracy and speed of aligning speech and text during training.

Mixup: Mixup is a common data augmentation strategy (Zhang et al., 2018; Jin et al., 2025). In E2E-ST, it is applied for bridging the modality gap (Fang et al., 2022; Zhou et al., 2023), where the model is trained on a discrete mixup of speech and text representations. Mixup training in E2E-ST requires an alignment between speech and text that can be generated using an alignment tool (Fang et al., 2022). Zhou et al. (2023) alleviate the need for an alignment tool by aligning speech and text representations using optimal transport. Our approach is similar to (Zhou et al., 2023), where we generate the alignments dynamically during training. However, instead of discretely mixing speech and text representations, we apply mixup as a linear interpolation.

DTW: DTW is an algorithm for measuring similarity between two sequences of varying length (Sakoe and Chiba, 1978). Due to this property, it has been widely applied to speech data (Juang, 1984; Furtuna, 2008; Muda et al., 2010), and also more specifically in the context of aligning speech and text sequences (i.e. forced alignment). For example, Aeneas (Pettarin, 2017) aligns speech and text utterances by transforming the text utterances into speech, then uses DTW to align the synthetic and original speech sequences. Kürzinger et al. (2020) uses an algorithm that resembles DTW by using dynamic programming and backtracking to find the optimal alignment based on Connectionist Temporal Classification (CTC) probabilities. In this work, we adapt DTW to dynamically align speech and text based on their embeddings.

3 Method

3.1 Architecture

Inspired by previous work in E2E-ST (Fang et al., 2022; Zhou et al., 2023), our model consists of two main components, a speech encoder, and a translation encoder-decoder. The translation encoder-decoder is a standard transformer model that can be decomposed into 3 components: a text embedding layer, an encoder that inputs either speech or text embeddings, and a decoder that generates the target sentence.

3.2 DTW for Aligning Speech and Text Representations

DTW can be used to compute similarities between two sequences of variable length along time. This is achieved by finding an optimal path between the two sequences, or the path that leads to their maximum similarity. The time dimension of the two sequences is said to be *warped*. In our case, when aligning speech and token embeddings, we only warp the token time dimension to have a one-to-many relationship from speech to token embeddings. We start by computing the cosine similarity between each speech embedding $t \in [0; N-1]$ to each token $j \in [0; M-1]$, then we use the similarity matrix $S \in \mathbb{R}^{N \times M}$ to compute a trellis matrix T of the same dimension:

$$T_{t,j} = \begin{cases} S_{t,j} & t = 0, j = 0 \\ -\infty & t = 0, j > 0 \\ +\infty & t > N - M, j = 0 \\ S_{t,j} + T_{t-1,j} & t > 0, j = 0 \\ \max(T_{t-1,j}, T_{t-1,j-1}) & \\ \quad + S_{t,j} & t > 0, j > 0 \end{cases} \quad (1)$$

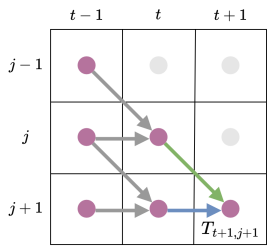


Figure 2: An illustration of the possible alignment paths. Each frame is assigned to only one token, while a token can be assigned to multiple frames.

The last step is backtracking, where we traverse the trellis starting from the last frame and token to find the optimal path, or the path with maximum similarity, which eventually represents the alignment a from speech to text tokens. We assign the last token to the last frame $a_{N-1} = M - 1$, and we traverse as follows:

$$a_t = \begin{cases} M-1 & t = N-1 \\ a_{t+1} - 1 & T_{t,a_{t+1}-1} > T_{t,a_{t+1}} \\ a_{t+1} & \text{else} \end{cases} \quad (2)$$

Figure 2 shows an illustration of the possible alignment paths. We can see that when backtracking from $t+1, j+1$ we can either go to the previous token j if $T_{t,j} > T_{t,j+1}$ or stay on token $j+1$ otherwise, which guarantees monotonicity. The constraints in the trellis matrix guarantee that all tokens are aligned to at least one frame, since the diagonal is filled with $-\infty$ during the trellis computation, when backtracking, this guarantees that we move to $j-1$ when $j \geq t$.

By fully vectorizing both the trellis computation and backtracking, our implementation achieves a much faster alignment.

3.3 Mixup Training

Given a sequence of speech representations generated using the speech encoder $f = [f_0, f_1, ..., f_{N-1}]$ and a sequence of text embeddings generated using the text embedding layer $e = [e_0, e_1, ..., e_{M-1}]$, our method generates an alignment $a = [a_0, a_1, ..., a_{N-1}]$ as described in §3.2. Finally, we apply mixup similarly to previous work (Zhou et al., 2023):

$$m_i = \begin{cases} f_i & p > p^* \\ e_{a_i} & else \end{cases} \quad (3)$$

where p^* is the mixup probability which controls how much text embeddings we introduce into the speech manifold, and p is sampled from a uniform distribution $\mathcal{U}(0, 1)$. We term this discrete mixup. We further introduce interpolation mixup (Zhang et al., 2018), where instead of selecting a speech or text embedding based on probability p^*, we use p^* as a mixup coefficient to linearly interpolate speech and text embeddings:

$$m_i = (1 - p^*).f_i + p^*.e_{a_i} \quad (4)$$

We argue that interpolation mixup can be more robust to alignment noise since the speech embeddings are not entirely replaced as in discrete mixup, but they are softly down-weighted. Therefore, even in the presence of alignment noise, the model still has access to the correct speech embeddings. Furthermore, it can be more data efficient, since all the speech and text token embeddings are included in training, rather than selecting one or the other.

3.4 Training Objective

We train with similar training objectives as CMOT (Zhou et al., 2023) to ensure fair comparison. The ST training corpus is denoted as $D = (s, x, y)$, where s is the speech input, x is the transcription, and y is the translation. In the first stage, the translation encoder-decoder is pre-trained on transcription-translation pairs using cross entropy:

$$\mathcal{L}_{MT} = -\mathbb{E}_{x,y} \log P(y|x) \quad (5)$$

The second stage is multi-task fine-tuning with ST and MT tasks using cross entropy:

$$\begin{aligned} \mathcal{L}_{ST} &= -\mathbb{E}_{s,x,y} \log P(y|s) \\ \mathcal{L}_{MT} &= -\mathbb{E}_{s,x,y} \log P(y|x) \end{aligned} \quad (6)$$

Furthermore, to bridge the modality gap between speech and text representations, we train with Kullback-Leibler (KL) divergence between the output probability distribution under mixup input m, and the output distribution of the ST task, as well as with the output distribution of the MT task:

$$\begin{aligned} \mathcal{L}_{KL_{m \leftrightarrow s}} = \mathbb{D}_{KL}(P(y|s)||P(y|m)) + \\ \mathbb{D}_{KL}(P(y|m)||P(y|s)) \end{aligned} \quad (7)$$

$$\begin{aligned} \mathcal{L}_{KL_{m \leftrightarrow x}} = \mathbb{D}_{KL}(P(y|x)||P(y|m)) + \\ \mathbb{D}_{KL}(P(y|m)||P(y|x)) \end{aligned} \quad (8)$$

Therefore, the final loss is:

$$\mathcal{L} = \mathcal{L}_{ST} + \mathcal{L}_{MT} + \lambda.(\mathcal{L}_{KL_{s \leftrightarrow m}} + \mathcal{L}_{KL_{x \leftrightarrow m}})/2 \quad (9)$$

where λ is a hyperparameter weight to control the KL losses.

4 Experiments

4.1 Dataset

We conduct our experiments on CoVoST-2 dataset (Wang et al., 2020), a large multilingual ST dataset that is based on Common Voice project (Ardila et al., 2020). CoVoST-2 covers translation from 21 source languages to English and from English to 15 target languages, and it contains speech, transcription and translation triplets. In this work, due to computational resources, we focus on 6 language directions: En-De, En-Ca, En-Ar, De-En, Fr-En, and Es-En. These directions are selected to ensure a balanced number of En-X and X-En directions. Furthermore, all languages selected are high resource with a minimum of 97 hours of training data and are of varying linguistic distance from English.

4.2 Experimental Setup

Pre-processing:
For speech input, we use the raw 16 bit 16kHz mono-channel audio. We filter out examples with a number of frames higher than 480k or less than 1k. For the text input, we remove punctuation, then we tokenize using a uni-gram SentencePiece model (Kudo and Richardson, 2018) with a vocabulary of 10k that is shared between the source and target languages.

Model:
Our model is composed of a speech encoder and a translation encoder-decoder. For the speech encoder, we use a pre-trained base HuBERT model (Hsu et al., 2021) for En-X language directions, and mHuBERT-147 (Zanon Boito et al., 2024) (a multilingual version of HuBERT base model) for X-En language directions. To shrink the audio representations over the time axis, we stack 2 1-dimensional convolution layers of kernel size 5, stride size 2, padding 2, and hidden dimension 1024. For the translation encoder, we use 6 transformer encoder layers. For the translation decoder, we use 6 transformer decoder layers. Each transformer layer is comprised of 512 hidden units, 8 attention heads, and 2048 feed-forward hidden units.

Training:
We train our model in two stages, first we pre-train the translation encoder-decoder on CoVoST2 transcription-translation pairs. We train with a learning rate of 1e-4, a maximum of 33k tokens per batch, and for a maximum 100k steps. We early stop training if the loss doesn't decrease for 10 epochs. During the second stage, we fine-tune the speech encoder and translation encoder-decoder with a learning rate of 1e-4, a maximum of 16M audio frames per batch, and we train for 40k steps. For CMOT, NFA-Align and DTW-Align, we train with a mixup probability $p^* = 0.2$ and a KL weight $\lambda = 2.0$.

The MT models are trained using one A100 GPU and ST models are trained using one H100 GPU. We use Fairseq [2] (Ott et al., 2019) for the implementation.

Evaluation:
We average the last 10 epoch checkpoints for evaluation, and generate with a beam size of 5. We use SacreBLEU (Post, 2018) to compute detokenized case-sensitive BLEU score (Papineni et al., 2002). We also use SacreBLEU to measure statistical significance using paired approximate randomization (Riezler and Maxwell, 2005).

Low Resource Setting:
All the languages in our experiments are considered high resource with at least 97 hours of training data, therefore, to evaluate our method in a low resource setting, we simulate a low resource scenario by sampling 10 hours of ST training data and 1 hour of development data for each language directions. During training, we use the same hyperparameters but we early stop if the loss did not decrease on the development set for 10 epochs. Our goal is to demonstrate how noise in the alignment has a more pronounced effect in low-resource ST scenarios. Therefore, we use a simulated low-resource setting with the same languages and training setup to avoid any confounding effects that would arise from using a different dataset.

4.3 Main Results

Baselines:
We experiment with the following models:
HuBERT-Transformer: Composed of speech encoder and translation encoder-decoder trained for ST.
CMOT: HuBERT-Transformer trained by using CMOT alignment for discrete mixup training.
NFA-Align: Using word level alignments from NeMo Forced Aligner (NFA) [3] which was shown to achieve state-of-the-art results in terms of alignment accuracy (Rastorgueva et al., 2023) for mixup training.
DTW-Align-Discrete (Ours): Using DTW for

[2] https://github.com/facebookresearch/fairseq
[3] https://github.com/NVIDIA/NeMo/tree/main/tools/nemo_forced_aligner

Model	En-De	En-Ca	En-Ar	De-En	Fr-En	Es-En	Avg.
Revist-ST ((Zhang et al., 2022))†	17.5	22.9	12.3	14.1	26.9	15.7	-
U2TT (Large) (Zhang et al., 2023a)†	-	-	-	16.7	27.4	28.1	-
DUB (Large) (Zhang et al., 2023a)†	-	-	-	19.5	29.5	**30.9**	-
SRPSE (Zhang et al., 2025)†	-	-	-	21.4	29.3	-	-
CoVoST-2 (Wang et al., 2020)†	18.4	23.6	13.9	18.9	27.0	28.0	21.6
CTC+OT (Le et al., 2023)†	20.6	26.5	15.3	20.4	28.4	29.2	23.4
HuBERT-Transformer	21.4	27.4	15.7	21.8	28.4	28.0	23.8
CMOT	**21.8**	28.2	**16.2**	23.6	30.9	29.6	**25.0**
NFA-Align	21.4	28.1	16.0	23.5	30.9	29.8	24.9
DTW-Align-Discrete	21.7	**28.3**	**16.2**	23.5	**31.0**	29.4	**25.0**
DTW-Align	**21.8**	28.2	16.1	**23.7**	30.8	29.5	**25.0**

Table 1: BLEU score results on CoVoST2 test set. The table shows that CMOT, DTW-Align, and DTW-Align-Discrete achieve the best results against other baselines. †: indicates results reported in the original work (the rest of the baselines are trained in this study)

generating alignments and training with discrete mixup (Equation 3) similar to CMOT.

DTW-Align (Ours): Using DTW for generating alignments and training with interpolation mixup (Equation 4).

5 Results and Discussion

Table 1 shows the results of our method and baselines, and results of previous work that was evaluated on the CoVoST2 dataset. The results show that consistent with previous studies (Fang et al., 2022), the baseline HuBERT-Transformer remains a competitive baseline, even outperforming previous work that uses more complex techniques. Furthermore, CMOT, DTW-Align-Discrete and DTW-Align achieve the best results overall. Although we train under similar settings and we do not optimize our method differently, we achieve similar results to CMOT. Surprisingly, NFA-Align which uses NFA to align speech and text lags slightly behind on average (i.e. 0.1 BLEU), this suggests that in a high resource setting, and with a low mixup probability the effect of noise in the alignment is less evident.

5.1 Alignment Accuracy and Training time

Table 2 shows that our method produces more accurate alignments with a significant increase of 19% in alignment accuracy. Furthermore, our method is more than 33 times faster in terms of execution time, which is concretely manifested in the staggering difference in training time between CMOT and DTW-Align (i.e. 14:20:53 and 6:48:14 respectively). As a reference, HuBERT-Transformer baseline training time is 6:32:53, which means that our method improves the performance over this baseline (by an average of 1.2 BLEU points) without the drawbacks of the significant training time overhead that CMOT suffers from. Therefore, although our method achieves similar results to state-of-the-art CMOT in high resource settings, it offers a significant advantage in terms of training time. In Section 6, we show that due to the improved alignment accuracy, our method is more robust both in low resource settings and under higher mixup probability values.

Method	Accuracy ↑	Execution Time ↓	Train Time ↓
CMOT	26%	97.89	14:20:53
DTW-Align	**45%**	**2.91**	**6:48:14**

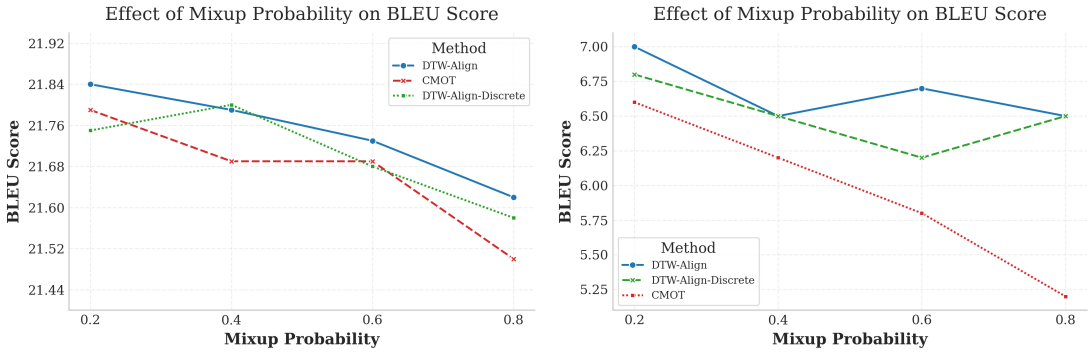
Table 2: We show the accuracy of alignments against NFA, and the execution time on CoVoST2 En-De dev set, plus the training time on En-De train set. DTW-Align is significantly faster and more accurate than CMOT.

6 Analysis

Although our method substantially outperfroms CMOT in terms of alignment accuracy, it does not yield improvements in ST performance. We attribute this to two factors: the amount of training data, which makes training more robust under noise and the low mixup probability value, which is set to 0.2. In §6.1 we measure the performance of CMOT, DTW-Align-Discrete and DTW-Align in a simulated low resource scenario of 10h per lan-

Model	En-De	En-Ca	En-Ar	De-En	Fr-En	Es-En	Avg.
HuBERT-Transformer	6.4	8.7	2.2	1.8	3.2	2.9	4.2
CMOT	6.6	9.6	2.7	2.8	8.5	7.5	6.3
DTW-Align-Discrete	6.8**	9.6	2.8	2.7	**8.6**	7.9**	6.4
DTW-Align	**7.0****	**9.8****	**3.1****	**2.9***	**8.6**	**8.0****	**6.6**

Table 3: BLEU score results on CoVoST2 test set in the low resource setting. The table shows that on overall DTW-Align-Discrete and DTW-Align on overall achieve better results than CMOT, with DTW-Align achieving the best results overall. *, ** indicate whether the improvement over CMOT is statistically significant with $p < 0.05$ and $p < 0.01$ respectively.



(a) En-De

(b) En-De Low Resource

Figure 3: The BLEU score of CMOT and DTW-Align under different mixup probabilities on En-De (Figure 3a) and En-De Low Resource (Figure 3b). DTW-Align is more robust to higher mixup probabilities than CMOT even with discrete mixup. This can be explained by noise in CMOT alignments.

guage direction, and in §6.2 we ablate the mixup probability value.

6.1 Low Resource Setting

Models can be more vulnerable to the negative effects of alignment noise in low resource scenarios. To study this, we compare the performance of CMOT, DTW-Align-Discrete and DTW-Align in a low resource setting of 10h of ST training data and 1h of development data. Table 3 shows the results over the 6 language directions in our experiments. Overall, DTW-Align-Discrete achieves better results than CMOT, with the improvements on En-De and Es-En being statistically significant. Furthermore, DTW-Align achieves the best results, with statistically significant improvement over CMOT on 5 language directions out of 6. These results show that combining the alignment accuracy of DTW and the robustness of interpolation mixup yields the best performance in low resource settings. Although our method performs on par with CMOT in high-resource settings, it offers an increase in performance in low resource ones, where effects of noise on CMOT are more pronounced. Finally, we find that the improvement of DTW-Align over HuBERT-Transformer has doubled (i.e. from 1.2 to 2.4 BLEU points), which demonstrates the advantage of mixup training in low resource settings.

6.2 Mixup Probability

We perform an ablation study on the effect of increasing the mixup probability p^* of CMOT, DTW-Align-Discrete and DTW-Align as shown in Figure 3 on En-De in high (Figure 3a) and low resource setting (Figure 3b). Results indicate that higher mixup probabilities lead to lower performance but the performance degradation is more significant in the case of CMOT, especially in the low resource setting, where training is more vulnerable to noise. This demonstrates that using DTW for aligning speech and text representations is more robust to the mixup probability hyperparameter, especially in low resource scenarios.

7 Conclusion

We introduce a method that eliminates the requirement for an external forced alignment tool by dynamically aligning speech and text embeddings

during training based on Dynamic Time Warping (DTW). Compared to state-of-the-art approaches, our method matches or exceeds BLEU score results while being significantly faster. We further demonstrate that using DTW-Align is more robust and data efficient in low resource settings. In addition, compared to HuBERT-Transformer baseline, our method improves performance by 1.2 and 2.4 BLEU points in high and low resource settings respectively with minimal overhead in the training time. Finally, unlike CMOT, our method can produce both token and word level alignments, which makes it compatible with previous work that requires word level alignments (Fang et al., 2022; Ouyang et al., 2023; Nguyen et al., 2025), therefore, it can bring a boost to the ongoing efforts on bridging the modality gap in E2E-ST or other speech-to-text tasks.

Limitations

Our work considers the following limitations:

Previous work shows that using external MT data for pretraining the translation encoder-decoder improves downstream ST performance. In our experiments, however, we only use internal CoVoST2 data for pretraining because of resource limitations.

Moreover, our work requires speech transcriptions, which might not be available for all languages. Future work can explore using transcriptions from an ASR model potentially extending the method's applicability to a wider range of languages.

Finally, CoVoST2 is an English centric dataset with English as the source or target language in all directions. Evaluating the accuracy and effect of speech and text alignment on other language directions would be valuable for future research.

Acknowledgments

The research presented in this paper was conducted as part of VOXReality project[4], which was funded by the European Union Horizon Europe program under grant agreement No 101070521.

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-11297.

[4] https://voxreality.eu/

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7050–7062, Dublin, Ireland. Association for Computational Linguistics.

Titus Furtuna. 2008. Dynamic programming algorithms in speech recognition. *Informatica Economica Journal*, XII.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2214–2225, Online. Association for Computational Linguistics.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Xin Jin, Hongyu Zhu, Siyuan Li, Zedong Wang, Zicheng Liu, Juanxi Tian, Chang Yu, Huafeng Qin, and Stan Z. Li. 2025. A survey on mixup augmentations and beyond. *Preprint*, arXiv:2409.05202.

B.-H. Juang. 1984. On the hidden markov model and dynamic time warping for speech recognition — a unified view. *AT&T Bell Laboratories Technical Journal*, 63(7):1213–1243.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. 2020. Ctc-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings*, page 267–278, Berlin, Heidelberg. Springer-Verlag.

Phuong-Hang Le, Hongyu Gong, Changhan Wang, Juan Pino, Benjamin Lecouteux, and Didier Schwab. 2023. Pre-training for speech translation: Ctc meets optimal transport. *Preprint*, arXiv:2301.11716.

Yuchen Liu, Hao Xiong, Jiajun Zhang, Zhongjun He, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. pages 1128–1132.

Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. 2010. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *J Comput*, 2.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. SpiRit-LM: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-aligned contrastive learning for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3891–3907, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alberto Pettarin. 2017. aeneas.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

E. Rastorgueva, V. Lavrukhin, and B. Ginsburg. 2023. Nemo forced aligner and its application to word alignment for subtitle generation. In *Proc. Interspeech 2023*, pages 5257–5258.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

H. Sakoe and S. Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4252–4261, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5099–5113, Seattle, United States. Association for Computational Linguistics.

Marcely Zanon Boito, Varun Iyer, Nathanaël Lagos, Laurent Besacier, and Ionut Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. In *Proc. Interspeech 2024*, pages 3939–3943.

Biao Zhang, Barry Haddow, and Rico Sennrich. 2022. Revisiting end-to-end speech-to-text translation from scratch. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26193–26205. PMLR.

Chengwei Zhang, Yue Zhou, Rui Zhao, Yidong Chen, and Xiaodong Shi. 2025. Representation purification for end-to-end speech translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6255–6269, Abu Dhabi, UAE. Association for Computational Linguistics.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023a. DUB: Discrete unit back-translation for speech translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7147–7164, Toronto, Canada. Association for Computational Linguistics.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Wei-Qiang Zhang. 2023b. Improving speech translation by cross-modal multi-grained contrastive learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:1075–1086.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Yuhao Zhang, Kaiqi Kou, Bei Li, Chen Xu, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2024. Soft alignment of modality space for end-to-end speech translation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11041–11045.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7873–7887, Toronto, Canada. Association for Computational Linguistics.

Targeted Source Text Editing for Machine Translation: Exploiting Quality Estimators and Large Language Models

Hyuga Koretaka[†*] **Atsushi Fujita**[‡] **Tomoyuki Kajiwara**[†]
[†]Graduate School of Science and Engineering, Ehime University, Ehime, Japan
[‡]National Institute of Information and Communications Technology (NICT), Kyoto, Japan
[†]{koretaka@ai., kajiwara@}cs.echime-u.ac.jp
[‡]atsushi.fujita@nict.go.jp

Abstract

To improve the translation quality of "black-box" machine translation (MT) systems, we focus on the automatic editing of source texts to be translated. In addition to the use of a large language model (LLM) to implement robust and accurate editing, we investigate the usefulness of targeted editing, i.e., instructing the LLM with a text span to be edited. Our method determines such source text spans using a span-level quality estimator, which identifies actual translation errors caused by the MT system of interest, and a word aligner, which identifies alignments between the tokens in the source text and translation hypothesis. Our empirical experiments with eight MT systems and ten test datasets for four translation directions confirmed the efficacy of our method in improving translation quality. Through analyses, we identified several characteristics of our method and that the segment-level quality estimator is a vital component of our method.

1 Introduction

In the last decade, the quality of machine translation (MT) outputs has been significantly improved as a result of the advancements of neural MT (NMT) and large language models (LLMs) and the accumulation of parallel data in the community. A number of new techniques for further improving translation quality, i.e., reducing translation errors, have been presented at conferences; however, proprietary MT services have tended to remain state of the art, presumably thanks to undisclosed technologies and massive in-house data. Such "black-box" systems are, in general, difficult to adapt for users' niche use cases in which texts with specific content domains or text styles are to be translated.

To obtain better translations using black-box MT systems, several strategies have been proposed. One such strategy is "pre-editing," i.e., editing given source texts to improve their translation by an MT system of interest. Although studies on automatic pre-editing have long been conducted (§2), two issues remain. First, existing methods have only limited editing ability. Various types of source text editing can potentially improve its translatability (Miyata and Fujita, 2017, 2021). However, in past studies, researchers have addressed only specific linguistic complexities or performed the regeneration of entire texts indiscriminately. Another issue is that researchers have performed pre-editing without referring to the actual translation generated by the MT system, despite the proven effectiveness of editing source text with reference to actual translation errors (Uchimoto et al., 2006; Resnik et al., 2010). Different MT systems have different error tendencies; thus, editing expressions that the MT system can translate well would result in new translation errors.

In this study, we automate "targeted source text editing" using (a) quality estimation models to determine what to edit on the basis of the translation errors caused by a target MT system and to search for a better translation and (b) LLMs to realize various types of editing as in Ki and Carpuat (2025). In our method (§3), a given source text is translated by the MT system, and errors in the output are identified by a span-level quality estimator. Then, using an LLM, our method edits the source text with a text span annotated as the source of the severest translation error. Guiding the editing process with a trigger error does not guarantee that the MT system can translate the edited text better (Miyata and Fujita, 2017, 2021). Thus, our method searches for a better translation by repeating text editing and MT, relying on a segment-level quality estimator.

Our empirical experiments with eight MT systems and ten test datasets for four translation directions confirmed the efficacy of our method in improving translation quality (§5). Our analyses

[*] This work was done during an internship of the first author at NICT.

Proceedings of the Tenth Conference on Machine Translation, pages 200–219
November 8-9, 2025

also revealed several characteristics of our method, including the diverse impact depending on the MT system and dataset, the necessity of improving the segment-level quality estimator, and controlled editing realized by tailored instruction (§6).

2 Previous Work

MT systems and services have gradually pervaded our lives, and have been incorporated into the human-centered translation production process adopted by translation/language service providers (ISO/TC37, 2017). Before they became sufficiently practical, researchers examined several approaches to human–MT interaction. Uchimoto et al. (2006) proposed the editing of source texts motivated by translation errors; the method was later named "targeted paraphrasing" by Resnik et al. (2010). Inspired by previous studies on targeted paraphrasing, Miyata and Fujita (2017, 2021) conducted manual investigations into the pre-editing strategy for exploiting black-box services based on statistical and neural MT. Through incrementally performing source text editing in four content domains referring to MT outputs, they found that most (80%–100%) of the source texts could eventually be edited so that they will lead to no translation errors and that human editors have performed diverse types of edits, not only limited to paraphrasing, that can improve translation quality.

Automatic "pre-editing" methods, which have been studied for three decades, can be classified into two groups. The first group focuses on specific linguistic phenomena that are difficult to translate, such as low frequency words, subject ellipsis, and long sentences, and avoids them relying on a set of rewriting rules based on morpho-syntactic information, corpus statistics, and neural language models (Shirai et al., 1993; Kim and Ehara, 1994; Yamaguchi et al., 1998; Shirai et al., 1998; Yoshimi, 2001; Mirkin et al., 2013; Štajner and Popovic, 2016; Štajner and Popović, 2018; Koretaka et al., 2023). However, each of these methods only covers a specific type of editing, among the diverse promising ones. Early methods are difficult to replicate for other source languages because of their heavy reliance on hand-crafted rules and resources.

Another line of research has attempted to regenerate entire source texts by regarding the task as monolingual translation and applying data-driven sequence-to-sequence decoding methods (Sun et al., 2010; Nanjo et al., 2012; Mirkin et al., 2013; Mehta et al., 2020). The performance of this approach is dominated by the characteristics and quantity of training parallel data. Since no parallel data have been specifically tailored for the purpose of pre-editing, except for small collections for manual analyses and evaluation, researchers have used monolingual parallel data that exhibit other monolingual tasks, such as text simplification and text revision, or synthetic parallel data generated by back-translating bilingual parallel data and round-trip translation of monolingual data. This approach has been proven effective for rule-based and statistical MT systems (Sun et al., 2010; Nanjo et al., 2012; Mirkin et al., 2013); in contrast, it does not necessarily work for NMT systems (Koretaka et al., 2023). Recently, Ki and Carpuat (2025) examined the utility of LLMs for source text editing. They compared several editing strategies and identified that the instruction for text simplification and selection based on quality estimation are effective.

Unlike manual investigations (Miyata and Fujita, 2017, 2021), both of the aforementioned groups of methods do not refer to the translation errors caused by the MT system of interest. Although a linguistically motivated pre-edit should be helpful in general, excessive editing of translatable expressions would introduce new translation errors.

3 Targeted Source Text Editing

Unlike existing "pre-editing" methods, we propose the editing of given source texts to avoid actual translation errors that the target MT system makes, i.e., the automation of the manual investigation process (Miyata and Fujita, 2017, 2021). Algorithm 1 shows the overall procedure of our method.

Given a source text src_0, our method first translates it using an MT system of interest T, and evaluates the quality of the generated hypothesis hyp_0 with reference to src_0 using a quality estimator Q. The pair of hyp_0 and its score initializes the best result (steps 1–3). The open list of errors to be edited is also initialized with translation errors in hyo_0 identified by an error detector E (steps 4–5).

Then, for pre-defined iterations N or until no errors remain (step 7), our method repeats the following steps: (a) identify the severest error and corresponding source text span (steps 8–9, §3.1), (b) edit the identified source span (step 10, §3.2), (c) translate the edited source text and evaluate the new hypothesis (steps 11–12, §3.3), and (d) search for the best translation (steps 8, 13–17, §3.4).

Algorithm 1: Proposed Error-Informed Source Text Editing Method.

Input : Original source text src_0, Translator T, Quality estimator Q, Error detector E, Maximum iteration N, Aligner A, Paraphraser P
Output : Best translation $best\_hyp$

1 $hyp_0 = T.translate(src_0)$
2 $best\_score = Q.evaluate(src_0, hyp_0)$
3 $best\_hyp = hyp_0$
4 $errs = E.detect\_errors(src_0, hyp_0)$
5 $cands = [\langle src_0, hyp_0, errs\rangle]$ `// open list`
6 $i = 1$
7 **while** $i < N \wedge cands \neq [\,]$ **do**
8 $\quad \langle src, hyp, err\rangle, cands = select\_one\_error(cands)$
9 $\quad src^{ann} = A.propagate\_error(src, hyp, err)$
10 $\quad src_i = P.paraphrase(src^{ann})$
11 $\quad hyp_i = T.translate(src_i)$
12 $\quad score_i = Q.evaluate(src_0, hyp_i)$
13 $\quad$ **if** $score_i > best\_score$ **then**
14 $\quad\quad best\_score = score_i$
15 $\quad\quad best\_hyp = hyp_i$
16 $\quad\quad errs = E.detect\_errors(src_i, hyp_i)$
17 $\quad\quad cands.append(\langle src_i, hyp_i, errs\rangle)$
18 $\quad i = i + 1$
19 **return** $best\_hyp$

3.1 Identification of the Source Text Span

In general, a translation hypothesis can contain multiple errors derived from dispersed source text spans. Following the incremental amelioration approach (Miyata and Fujita, 2017, 2021), we focus on the severest error and corresponding source text span (steps 8–9) in each iteration.

To this end, we rely on an error detector or a span-level quality estimator E, which identifies error spans with a severity score (steps 4 and 16). E may not jointly detect source text spans each corresponding to an error in the hypothesis. We thus identify such spans by aligning the source text tokens and those in the hypothesis using an aligner A (step 9). Then, we determine the source span that aligns with the severest error. Other tuples of an error span, its corresponding source text span, and its severity score are stored in the open list (steps 5 and 17) for future iterations (§3.4).

3.2 Targeted Text Editing Using an LLM

Given a source text annotated with a text span, our method then attempts to edit the span. Since the annotated span can be linguistically diverse, from a single symbol or word to the entire source text, we require a robust editor that can realize diverse types of text editing without introducing linguistic errors and semantic changes.

To perform such monolingual text editing, we use a decoder-only LLM, assuming that it has learned diverse linguistic phenomena from massive text data and is well-instructed for various text-editing tasks. In addition to the source text and the text span to be edited, it would be useful to instruct the LLM on the text editing task, such as its sub-steps and constraints, with some examples if possible. To better control its output, instructing the LLM with prompt formatting through few-shot examples is a promising approach (He et al., 2024). However, we need to prepare countermeasures against irregular outputs, such as control sequences and an off-target format.

3.3 Translation and Evaluation

The edited source text src_i is translated by the MT system T (step 11), and the quality of the generated hypothesis hyp_i is evaluated by the quality estimator Q (step 12). Text editing is not necessarily successful; it may fail to edit the annotated error source, thereby resulting in semantic drift in src_i and/or severer errors in hyp_i. Therefore, the quality is evaluated with respect to the original source text src_0 rather than the corresponding source text src_i as in Ki and Carpuat (2025).

3.4 Search for the Best Translation

We search for a better translation through repeating targeted source text editing, hypothesis generation with the MT system, and quality assessment. Given the high computational cost for LLM-based text editing, it is not feasible to traverse the entire search space. Therefore, we conduct depth-first search as in past manual investigations (Miyata and Fujita, 2017, 2021): our method performs source text editing greedily as long as quality improves. If an edit is confirmed to be detrimental by the quality estimator, it discards the edited text and selects the second severest error of the previous version of the source text. If no error remains, it backtracks to one more previous version of the source text. This is implemented in the "`select_one_error()`" function (step 8).

4 Preliminary Experiments

We determined the detailed settings using a small set of Japanese-to-English translation examples and one MT system. Henceforth, resources used in our experiments, including datasets, pre-trained model checkpoints, and tools, will be presented in this manner. See Appendix A for their details.

For this purpose, we used an NMT model pre-trained on JParaCrawl (Morishita et al., 2022) (the big model) and four sets of Japanese–English parallel data: (a) the development data of ASPEC (Nakazawa et al., 2016) (abstracts of scientific papers), (b) the test data of WMT22 (Kocmi et al., 2022) (mixture of several domains), (c) the test data of MTNT (Michel and Neubig, 2018) (users' posts on social media), and (d) the development data of the Kyoto Free Translation Task (benchmark splits of the Japanese–English Bilingual Corpus of Wikipedia's Kyoto Articles; henceforth, KFTT). First, we translated the Japanese side of these datasets into English using the MT model and Fairseq (Ott et al., 2019), and detected translation errors using a span-level quality estimator, XCOMET-XL (Guerreiro et al., 2024). We then randomly extracted 25 text pairs from each dataset that contained at least one "critical" or "major" error. Referring only to the sampled 100 ($= 25 \times 4$) text pairs, we determined the details of our method, including the combination of the span-level quality estimator and word aligner (Appendix C.1), the LLM for source text editing (Appendix C.2), and the prompt template (Appendix B.1).

To identify text spans in the source texts to be edited, our method propagates errors identified by XCOMET-XL, relying on optimal transport implemented by OTAlign (Arase et al., 2023). We determined the alignments between the tokens in the source text and hypothesis using contextual token embeddings obtained by InfoXLM-Base (Chi et al., 2021), uniform distribution as the mass for each token, cosine distance between token embeddings as the cost function, Sinkhorn algorithm, and 0.1 as the weight for the entropy-based regularizer.

For source text editing, we used Llama-3.1-Swallow-70B-Instruct-v0.1 (henceforth, Llama-Swallow) (Fujii et al., 2024), an LLM trained on massive text data of the source language, i.e., Japanese, and devised a prompt template for it, including the way of specifying the source text span to be edited and providing a one-shot paraphrase example. Such an example was randomly sampled from the 71 paraphrase examples in Japanese taken from a taxonomy of paraphrases.[1] The instruction was formatted by applying a chat template using the Language Model Evaluation Harness.

To select the best translation hypothesis among those derived from different versions of source texts, we also used XCOMET-XL, a reference-free quality estimation metric.

Through the preliminary experiments, we observed that the translation quality achieved by our method saturated up to five iterations (Appendix C). This is fewer than the 5.4–21.8 iterations required to obtain an acceptable translation in a manual investigation (Miyata and Fujita, 2021) because of the limited ability of our method compared with humans. The advancement of each component, as well as their tight integration, should improve the ability of our method.

5 Evaluation

To confirm the efficacy of our method on translation quality, we conducted experiments. Although most components of our method are multilingual, we had only Japanese and English speakers to write prompt templates for source text editing and translation with LLMs. Therefore, we evaluated the applicability of our method on Japanese-to-English (Ja→En), Japanese-to-Chinese (Ja→Zh), English-to-Japanese (En→Ja), and English-to-Chinese (En→Zh) translation directions.

5.1 Settings

5.1.1 Configuration of the Proposed Method

The configuration of our method mostly follows the details that we determined in our preliminary experiments (§4). For both translation error detection and translation quality estimation, we used XCOMET-XL, with one exception: we regarded text spans annotated as "critical," "major," or "minor" as errors, thereby extending the target. We used OTAlign and InfoXLM-Base for propagating erroneous spans to the source text.

Source text editing in our method is a monolingual task. Considering that an LLM trained specifically for the source language should perform better than other LLMs (Appendix C.2), we used Llama-Swallow for the Ja→En and Ja→Zh tasks, and Llama-3.1-70B-Instruct (henceforth, Llama) (Grattafiori et al., 2024) for the En→Ja and En→Zh tasks. We used 71 Japanese and 44 English examples available in the aforementioned paraphrase taxonomy[1] as the pool for the one-shot demonstration. The prompt templates are shown in Appendix B.1.

We set the number of maximum iterations, i.e., N in Algorithm 1, to 5, following our preliminary experiments (§4).

[1] https://paraphrasing.org/paraphrase.html

5.1.2 MT Systems

We applied our method to eight MT systems: four NMT and four LLM-based systems. Although we chose publicly available checkpoints for the sake of reproducibility, we regarded them as black boxes with the aim of simulating applications of our method to proprietary MT systems and services.

The NMT systems were NLLB-200-3B (henceforth, NLLB) (NLLB Team et al., 2022) and three sized-variants of the Ja→En and En→Ja specific models trained on JParaCrawl (Morishita et al., 2022), labeled as small, base, and big.

The four LLM-based MT systems were based on two LLMs, i.e., Llama and Llama-Swallow (see Appendix B.2 for their prompt templates), and two methods for selecting a translation example from a reference parallel corpus. One is BM25 (Robertson and Zaragoza, 2009), implemented in bm25s (Lù, 2024), which searches the parallel corpus for a text pair whose source side is most similar to the given source text. To this end, the source text to be translated and the corresponding side of the parallel corpora were tokenized with MeCab (Kudo et al., 2004) and Moses tokenizer (Koehn et al., 2007) for Japanese and English, respectively. The other method, called "vector," identifies such a text pair relying on sentence embeddings. We used LaBSE (Feng et al., 2022) as the sentence encoder and Faiss (Douze et al., 2024) for search, where we indexed the source side of the parallel corpora using product quantization with the number of subquantizers of 96 and the number of bits per subvector index of 8. As the reference parallel corpora, we used the official training data of WMT23 (Kocmi et al., 2023) consisting of 33.9M and 55.2M text pairs for Japanese–English and Chinese–English pairs, and Japanese–Chinese JParaCrawl (Nagata et al., 2024) consisting of 4.6M text pairs.

5.1.3 Test Datasets

For Ja→En, we used four datasets: [a] the test data of ASPEC (Nakazawa et al., 2016), [b] the test data of WMT23 (Kocmi et al., 2023) (mixture of several domains), [c] the test data of MTNT19 (Li et al., 2019) (users' posts on social media), and [d] the test data of KFTT. For Ja→Zh, we used [e] the test data of ASPEC. For En→Ja, we used four datasets: [f] the test data of the Asian Language Treebank (Riza et al., 2016) (news articles; henceforth, ALT), [g] the test data of WMT23, [h] the test data of MTNT19, and [i] the test data of IWSLT 2017 (Cettolo et al., 2017) (TED talks; henceforth, IWSLT). We also used [j] the test data of IWSLT 2017 for En→Zh.

When these datasets were translated by the target MT systems, 61.6%–96.2% of the resulting translations contained at least one "critical," "major," or "minor" error (see Appendix D for the details). Note that our method processes only these "erroneous test subsets."

5.1.4 Other Methods Compared

We regarded translation of the original source texts, i.e., hyp_0 in Algorithm 1, as the baseline. In addition, we evaluated the following "non-targeted" methods. Unlike ours, they are unaware of the target MT system and attempt to pre-edit source texts irrespective of potential translation errors.

Word-Sub: We replicated the word-substitution method proposed by Koretaka et al. (2023), which generates N-best paraphrases by substituting one word, relying on a masked language model and cosine similarity between word embeddings. We used language-specific BERT models trained through whole-word masking (Tohoku-NLP BERT base Japanese and Google BERT large for English) and FastText word embeddings.

Seq2seq-B: Although it is proven ineffective (Koretaka et al., 2023), we trained a monolingual sequence-to-sequence model for each source language, following Wieting et al. (2017) (see Appendix G for the training details), and obtained N paraphrased texts using the model via beam search with a beam size of 12.[2]

LLM-NT (non-targeted): We obtained N versions of source texts through iterative paraphrasing with an LLM.[3] The only difference from our method is that the source texts were not annotated on the basis of translation errors. We thus derived the prompt templates for Llama-Swallow and Llama from those for our method (Appendix B.1) by removing the step-wise instruction for error-informed text editing while retaining the criteria to meet.

Each of the N paraphrased source texts and the original source text was translated separately by the given MT system, and the best translation among

[2]We also examined nucleus sampling with $top\_p = 0.95$ but it underperformed beam search in most configurations.

[3]With $N = 1$, this is similar to one of the MT-Agnostic rewriting methods examined by Ki and Carpuat (2025).

MT System	Editing Method	Ja→En				Ja→Zh	En→Ja				En→Zh	#+	#-	#w
		ASPEC [a]	WMT23 [b]	MTNT19 [c]	KFTT [d]	ASPEC [e]	ALT [f]	WMT23 [g]	MTNT19 [h]	IWSLT [i]	IWSLT [j]			
NLLB [1]	Baseline	81.18	74.80	68.59	64.28	83.76	87.99	81.41	77.23	82.83	79.51	-	-	0
	Word-Sub	**81.47***	**76.40***	**70.50***	**66.91***	**84.05***	**88.70***	**83.35***	**79.43***	**84.44***	**79.95***	10	0	1
	Seq2seq-B	81.08	**75.78***	**70.09***	**66.51***	83.55*	**88.86***	**83.29***	**79.25***	**84.16***	**79.97***	8	1	1
	LLM-NT	**81.58***	**76.73***	**71.32***	**66.52***	**84.10***	**88.85***	**84.56***	**81.35***	**85.09***	**80.68***	10	0	7
	Ours	**81.67***	**76.24***	**70.36***	**65.90***	**83.90***	**88.52***	**82.71***	**78.50***	**83.29***	**79.83***	10	0	1
JParaCrawl (small) [2]	Baseline	81.75	76.92	72.77	73.46	-	85.74	80.26	74.69	79.73	-	-	-	0
	Word-Sub	**81.95***	**77.85***	**73.44***	**73.68**	-	**86.44***	**80.51**	74.04*	**80.63***	-	5	1	0
	Seq2seq-B	81.66	**77.29***	72.72	73.14	-	**86.60***	**80.84***	**74.73**	**80.42***	-	4	0	0
	LLM-NT	**82.12***	**78.25***	**74.17***	**74.11***	-	**87.15***	**83.38***	**79.36***	**82.74***	-	8	0	5
	Ours	**82.27***	**78.57***	**74.08***	**74.19***	-	**86.74***	**82.34***	**77.29***	**81.88***	-	8	0	3
JParaCrawl (base) [3]	Baseline	82.30	77.75	72.67	74.52	-	86.77	80.96	75.22	80.43	-	-	-	0
	Word-Sub	**82.45***	**78.40***	**73.76***	**74.63**	-	**87.29***	**81.35***	73.77*	**81.13***	-	6	1	0
	Seq2seq-B	82.09*	**78.00***	**72.85**	74.32	-	**87.19***	**81.64***	**75.26**	**81.00***	-	4	1	0
	LLM-NT	**82.57***	**78.98***	**74.54***	**75.09***	-	**87.93***	**83.84***	**79.66***	**83.15***	-	8	0	4
	Ours	**82.71***	**79.15***	**74.49***	**75.30***	-	**87.94***	**83.14***	**77.97***	**82.58***	-	8	0	4
JParaCrawl (big) [4]	Baseline	82.87	79.26	74.96	76.25	-	88.04	82.31	76.36	81.63	-	-	-	0
	Word-Sub	**82.96**	**79.61***	74.23*	**76.50**	-	**88.30***	**82.78***	75.70*	**82.21***	-	4	2	0
	Seq2seq-B	82.63*	79.26	74.35*	75.27*	-	**88.45***	**82.75***	**76.55**	**82.22***	-	3	3	0
	LLM-NT	**83.09***	**79.96***	**75.68***	**76.76***	-	**88.94***	**84.75***	**80.59***	**83.96***	-	8	0	6
	Ours	**83.06***	**80.01***	**75.72***	**76.50**	-	**88.91***	**84.12***	**79.32***	**83.19***	-	7	0	2
Llama (BM25) [5]	Baseline	81.61	80.45	75.26	76.96	86.57	89.62	85.51	82.32	81.64	81.86	-	-	0
	Word-Sub	**82.92***	**81.20***	**76.17***	**77.98***	**86.72**	**89.96***	**86.25***	**82.88***	**83.13***	**83.00***	9	0	6
	Seq2sec-B	**82.63***	**80.71**	**75.84***	**77.46**	86.34*	**89.96***	**86.23***	**83.04***	**82.91***	**83.07***	7	1	1
	LLM-NT	**82.56***	**81.13***	**76.32***	**77.71***	**86.69***	**89.63**	**86.12***	**83.31***	**83.89***	**83.17***	9	0	4
	Ours	**81.76***	**80.98***	**75.93***	**77.41***	86.57	**89.70**	**86.06***	**83.25***	**83.33***	**82.79***	8	0	0
Llama (vector) [6]	Baseline	81.98	80.67	74.75	76.82	86.37	89.86	85.92	82.44	81.85	82.21	-	-	0
	Word-Sub	**82.90***	**81.47***	**76.07***	**77.96***	**86.81***	**89.92**	**86.54***	**83.49***	**83.57***	**83.05***	9	0	6
	Seq2sec-B	**82.57***	**80.82**	**75.55***	**77.36***	86.31	**90.00**	**86.56***	**82.76**	**82.91***	**82.74***	6	0	2
	LLM-NT	**82.57***	**81.23***	**75.80***	**77.44***	**86.65***	89.70	**86.21**	**83.74***	**84.16***	**82.94***	8	0	2
	Ours	**82.16***	**81.08***	**75.54***	**77.00**	**86.38**	**90.00**	**86.38***	**83.10***	**83.10***	**82.78***	7	0	1
Llama-Swallow (BM25) [7]	Baseline	80.84	80.52	75.14	77.15	86.54	90.68	86.40	83.38	83.83	81.83	-	-	1
	Word-Sub	**82.06***	**81.60***	**76.32***	**78.45***	86.37*	**90.96***	**87.12***	**84.08***	**85.34***	**82.20***	9	1	2
	Seq2sec-B	**82.41***	**81.25***	**76.40***	**77.97***	86.02*	**90.92***	**87.01***	**83.93***	**85.04***	**82.47***	9	1	1
	LLM-NT	**82.08***	**81.66***	**76.44***	**78.09***	86.51	**90.80**	**87.16***	**84.38***	**85.69***	**82.52***	8	0	6
	Ours	80.76	**80.85***	**75.27**	**77.19**	86.20*	**90.92***	**87.01***	**84.05***	**85.14***	**82.21***	6	1	0
Llama-Swallow (vector) [8]	Baseline	81.40	80.93	74.62	77.69	86.48	90.85	86.75	83.46	84.34	81.76	-	-	1
	Word-Sub	**82.64***	**81.94***	**76.38***	**78.38***	86.36	**91.16***	**87.47***	**84.06***	**85.67***	**82.48***	9	0	7
	Seq2sec-B	**82.72***	**81.21**	**75.87***	**78.00**	86.05*	**91.05***	**87.33***	**83.96***	**85.05***	**82.30***	7	1	1
	LLM-NT	**82.36***	**81.65***	**76.27***	**78.25***	86.37	**90.96**	**87.22***	**84.29***	**85.62***	**82.46***	8	0	0
	Ours	81.38	**81.33***	**75.32***	77.58	86.07*	**90.96**	**87.16***	**84.30***	**85.43***	**82.01**	5	1	1

Table 1: COMET scores for the entire test sets. **Bold** indicates the improvement over the Baseline, underlining so does the best score among all the methods, the "#+" and "#-" columns show the number of test datasets for which the method achieved a significantly better or worse score (marked with "*") than the Baseline, respectively, and the "#w" column presents the number of datasets for which the method achieved the best score for each MT system.

$(N+1)$ hypotheses was selected by XCOMET-XL similarly to our method.[4]

5.1.5 Evaluation Metric

To evaluate the translation quality of MT outputs, we used the COMET score (Rei et al., 2020), specifically with the wmt22-comet-da checkpoint (Rei et al., 2022). We performed paired bootstrap resampling (Koehn, 2004) to test the statistical significance of the score difference from the baseline.

[4]XCOMET-XL achieved consistently better results than mBART used in Koretaka et al. (2023) in our preliminary experiments, although it was substantially slow.

5.2 Results

The COMET scores of all the methods in 74 test configurations are presented in Table 1. For the four NMT systems (the upper half), the LLM-based text editing methods, i.e., LLM-NT and ours, significantly improved translation quality, with only one exception. The word-substitution method also led to a significant gain in roughly 70% of configurations, outperforming the sequence-to-sequence method. For the LLM-based MT systems (the bottom half), the LLM-based text editing methods and the word-substitution method also achieved significant improvements. The word-substitution method

MT System	Editing Method	Ja→En				Ja→Zh	En→Ja				En→Zh	#+	#-	#w
		ASPEC [a]	WMT23 [b]	MTNT19 [c]	KFTT [d]	ASPEC [e]	ALT [f]	WMT23 [g]	MTNT19 [h]	IWSLT [i]	IWSLT [j]			
NLLB [1]	Baseline	79.71	72.94	67.92	66.49	84.01	87.48	82.25	78.79	83.72	79.70	-	-	0
	Word-Sub	**80.07***	**74.28***	**69.34***	**68.20***	**84.16***	**87.95***	**83.22***	**79.64***	**84.12***	79.69	9	0	0
	Seq2seq-B	79.48*	**73.62***	**68.70***	**67.54***	83.64*	**88.11***	**83.19***	**79.66***	**84.13***	**79.89**	7	2	0
	LLM-NT	**80.21***	**74.96***	**70.26***	**68.01***	**84.25***	**88.13***	**84.22***	**81.20***	**84.55***	**80.26***	10	0	5
	Ours	**80.43***	**75.11***	**70.48***	**68.44***	**84.16***	**88.15***	**83.90***	**80.44***	**84.33***	**80.09***	10	0	5
JParaCrawl (small) [2]	Baseline	80.38	74.70	71.54	73.45	-	84.98	79.25	73.80	78.18	-	-	-	0
	Word-Sub	**80.66***	**75.78***	**72.13***	**73.61**	-	**85.76***	**79.53**	73.05*	**79.18***	-	5	1	0
	Seq2seq-B	80.25	**75.10***	71.30	73.10*	-	**85.89***	**79.87***	73.77	**78.92***	-	4	1	0
	LLM-NT	**80.86***	**76.43***	**72.80***	**74.06***	-	**86.45***	**82.65***	**78.71***	**81.66***	-	8	0	4
	Ours	**81.10***	**76.82***	**73.16***	**74.24***	-	**86.11***	**81.62***	**76.74***	**80.76***	-	8	0	4
JParaCrawl (base) [3]	Baseline	81.04	75.38	70.84	74.48	-	85.91	79.87	74.51	78.93	-	-	-	0
	Word-Sub	**81.24***	**76.18***	**72.08***	**74.56**	-	**86.46***	**80.23***	72.98*	**79.72***	-	6	1	0
	Seq2seq-B	80.76*	**75.63**	**70.88**	74.25	-	**86.39***	**80.56***	74.51	**79.61***	-	3	1	0
	LLM-NT	**81.40***	**76.96***	**72.92***	**75.09***	-	**87.22***	**83.04***	**79.19***	**82.11***	-	8	0	3
	Ours	**81.60***	**77.21***	**73.12***	**75.32***	-	**87.27***	**82.37***	**77.66***	**81.51***	-	8	0	5
JParaCrawl (big) [4]	Baseline	81.61	77.10	73.20	76.34	-	87.08	81.00	75.62	79.82	-	-	-	0
	Word-Sub	**81.73**	**77.53***	72.17*	**76.58**	-	**87.38***	**81.50***	74.86*	**80.58***	-	4	2	0
	Seq2seq-B	81.27*	77.00	72.45*	75.41*	-	**87.56***	**81.49***	**75.76**	**80.56***	-	3	3	0
	LLM-NT	**81.91***	**78.02***	**73.90***	**76.82***	-	**88.14***	**83.76***	**80.11***	**82.76***	-	8	0	6
	Ours	**81.90***	**78.13***	**74.18***	**76.61**	-	**88.12***	**83.14***	**78.98***	**81.78***	-	7	0	2
Llama (BM25) [5]	Baseline	82.05	79.24	74.23	78.47	86.58	89.02	84.76	81.39	81.02	81.40	-	-	0
	Word-Sub	**82.41***	**79.69***	**74.51**	**78.85***	**86.62**	**89.25**	**85.32***	**81.81**	**82.34***	**82.29***	6	0	1
	Seq2seq-B	81.86*	78.80*	73.77	78.05*	86.21*	**89.35***	**85.37***	**81.93***	**82.11***	**82.38***	5	4	1
	LLM-NT	**82.25***	**79.70***	**74.84***	**78.58**	**86.65**	89.01	**85.21***	**82.29***	**83.18***	**82.44***	7	0	1
	Ours	**82.27***	**80.04***	**75.17***	**78.96***	86.57	**89.12**	**85.46***	**82.55***	**83.22***	**82.48***	8	0	7
Llama (vector) [6]	Baseline	82.10	79.48	73.75	78.25	86.61	89.23	84.89	81.52	81.22	81.66	-	-	0
	Word-Sub	**82.33***	**79.93***	**74.44***	**78.54**	**86.69**	**89.24**	**85.49***	**82.36***	**82.64***	**82.39***	7	0	2
	Seq2seq-B	81.77*	79.02*	**73.79**	77.75*	86.18*	**89.25**	**85.50***	**81.66**	**82.10***	**82.07***	3	4	1
	LLM-NT	**82.28***	**79.79***	**74.40***	**78.57**	**86.68**	88.91	**85.07**	**82.59***	**83.36***	**82.25***	4	1	3
	Ours	**82.36***	**80.10***	**74.87***	**78.45**	**86.62**	**89.43**	**85.48***	**82.36***	**82.87***	**82.32***	7	0	4
Llama-Swallow (BM25) [7]	Baseline	82.33	79.79	75.29	78.76	86.62	89.82	85.31	82.27	82.61	81.23	-	-	1
	Word-Sub	82.12*	**79.90**	74.90	**78.86**	86.29*	**90.20***	**86.07***	**82.75***	**84.30***	**81.50**	4	2	1
	Seq2seq-B	81.81*	79.39*	74.59*	78.30*	85.88*	**90.10***	**85.95***	**82.59**	**83.77***	**81.85***	4	5	1
	LLM-NT	**82.41**	**79.92**	75.08	**78.98**	86.49*	**89.94**	**86.14***	**83.10***	**84.63***	**81.78***	4	1	4
	Ours	82.20	**80.32***	**75.49**	**78.79**	86.25*	**90.15***	**86.13***	**83.14***	**84.40***	**81.66***	6	1	3
Llama-Swallow (vector) [8]	Baseline	82.29	80.03	74.50	78.85	86.59	90.00	85.75	82.13	83.18	81.20	-	-	1
	Word-Sub	82.24	**80.22**	**74.99**	78.78	86.29*	**90.32***	**86.46***	**82.73***	**84.64***	**81.81***	5	1	3
	Seq2seq-B	81.84*	79.40*	74.32	78.35*	85.93*	**90.24***	**86.29***	**82.59***	**83.88***	**81.66***	5	4	0
	LLM-NT	**82.37**	**80.28**	**75.23***	**78.97**	86.37*	**90.08**	**86.12***	**83.10***	**84.49***	**81.74***	5	1	2
	Ours	82.25	**80.67***	**75.55***	78.73	86.14*	**90.15**	**86.31***	**83.22***	**84.69***	**81.47**	5	1	4

Table 2: COMET scores for the erroneous test subsets (§5.1.3). See Table 1 for text decoration and symbols.

achieved the highest COMET score in more than 50% of configurations, whereas our method lagged behind it and LLM-NT.

Unlike existing non-targeted methods, our method edits only the erroneous test subsets (§5.1.3). Table 2 compares the COMET scores for these subsets, revealing the advantage of our method and diminished impact of the non-targeted methods. Our method achieved the highest COMET score in 34 out of 74 configurations, followed by LLM-NT which won in 28.

From these results, we conclude that the LLMs performed source text editing for MT more robustly and accurately than the existing methods. However, our targeted method is not yet incontestably superior over its non-targeted counterpart, i.e., LLM-NT. For instance, the prompt template developed with Ja→En examples had a minimal or negative impact on the Ja→Zh task, in particular with Llama-Swallow. In contrast, the equivalent prompt template manually translated into English worked fairly well, encouraging future applications of our method to other source languages.

6 Analyses

To better understand the characteristics of our method, we conducted several analyses, focusing on the erroneous test subsets (§5.1.3).

6.1 System-wise and Dataset-wise Impact

We hypothesize that the worse the quality of an MT output for the original source text is, the more

MT System	Ja→En				Ja→Zh	En→Ja				En→Zh
	ASPEC [a]	WMT23 [b]	MTNT19 [c]	KFTT [d]	ASPEC [e]	ALT [f]	WMT23 [g]	MTNT19 [h]	IWSLT [i]	IWSLT [j]
NLLB [1]	-0.381	-0.401	-0.416	-0.388	-0.324	-0.453	-0.540	-0.444	-0.340	-0.359
JParaCrawl (small) [2]	-0.368	-0.437	-0.339	-0.292	-	-0.375	-0.434	-0.340	-0.452	-
JParaCrawl (base) [3]	-0.392	-0.389	-0.385	-0.324	-	-0.493	-0.454	-0.373	-0.489	-
JParaCrawl (big) [4]	-0.393	-0.391	-0.382	-0.275	-	-0.479	-0.481	-0.437	-0.461	-
Llama (BM25) [5]	-0.353	-0.381	-0.322	-0.342	-0.269	-0.237	-0.400	-0.385	-0.633	-0.565
Llama (vector) [6]	-0.340	-0.380	-0.342	-0.231	-0.297	-0.264	-0.369	-0.347	-0.573	-0.486
Llama-Swallow (BM25) [7]	-0.228	-0.338	-0.335	-0.262	-0.215	-0.480	-0.466	-0.391	-0.683	-0.457
Llama-Swallow (vector) [8]	-0.281	-0.345	-0.347	-0.163	-0.157	-0.470	-0.394	-0.395	-0.619	-0.460

Table 3: Pearson product-moment correlation coefficients r between the baseline COMET score and its gain achieved by our method. See Appendix D for the number of segments in each configuration.

it should benefit from avoiding underlying translation errors by source text editing. To examine this, we calculated the correlation between the baseline COMET score and its gain. Table 3 summarizes segment-level correlation coefficients. Although there was moderate negative correlation for most configurations, we observed that others, such as the Ja→Zh ASPEC dataset translated by the LLM-based MT systems, did not follow the rule. In general, correlation for the LLM-based MT systems was weaker than those for the NMT systems, except for the two IWSLT tasks, and more diverse over the datasets. This implies that these LLMs had peculiar characteristics. For instance, some of the test datasets might have already been learned by them, unlike the NMT systems.

Figure 1 visualizes the correspondences between the baseline COMET score and its gain for each erroneous test subset. Our method had a stronger correlation than the other methods. One may consider that our method could be less impactful for very accurate MT systems, such as "black-box" proprietary systems. Despite this, we consider that our approach still has potential, because its advancement is orthogonal to the enhancements of those MT systems. For instance, we developed our method using only one NMT system and a small sample of Ja→En sentence pairs (§4), but it worked well for other translation directions (4f–4i) and stronger LLM-based MT systems (e.g., 5g–8i). Through our extensive experiments, we identified configurations where the current form of our method did not work well, such as the two ASPEC and KFTT tasks. We will conduct in-depth analyses to explore the reasons and address them in our future work.

6.2 Quality Estimator

We investigated whether the segment-level quality estimator, i.e., XCOMET-XL, was useful for search-

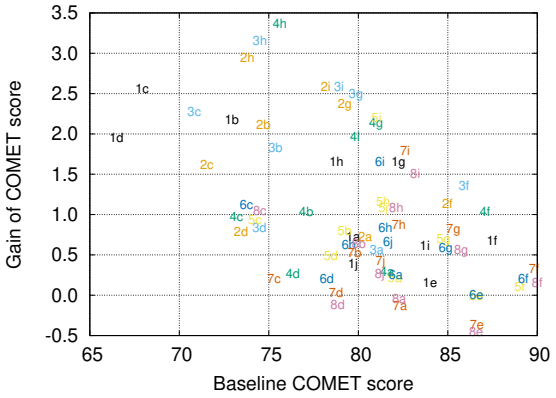


Figure 1: Baseline COMET score for the erroneous test subsets and its gain achieved by our method ($r = -0.522$). "1" to "8" and "a" to "j" are the indices of the MT systems and test datasets, respectively.

ing for the best translation. Figure 2 shows that the estimated quality monotonically improved during the search as intended; 1.91–8.48 and 2.46–11.05 points with $N = 1$ and $N = 5$, respectively. However, the final COMET score shown in Figure 3 did not follow the same trend, even though these two measures correlated well at the segment level in our experiments (0.292–0.762, Appendix E). For instance, in the two configurations where our method significantly deteriorated the COMET score, i.e., the two variants based on Llama-Swallow for the Ja→Zh ASPEC task (7e and 8e shown at top right of Figure 3), the correlation coefficient between the two measures was moderate (0.534 and 0.519). In contrast, the configurations with a weakest correlation, i.e., the JParaCrawl variants applied to the En→Ja IWSLT task (2e–4e, $r = 0.292$–0.296), achieved a 2.0–2.5 COMET point gain.

This discrepancy suggests that the segment-level quality estimator was a vital component, and thus requires further improvements to capture subtle differences between accurate translations.

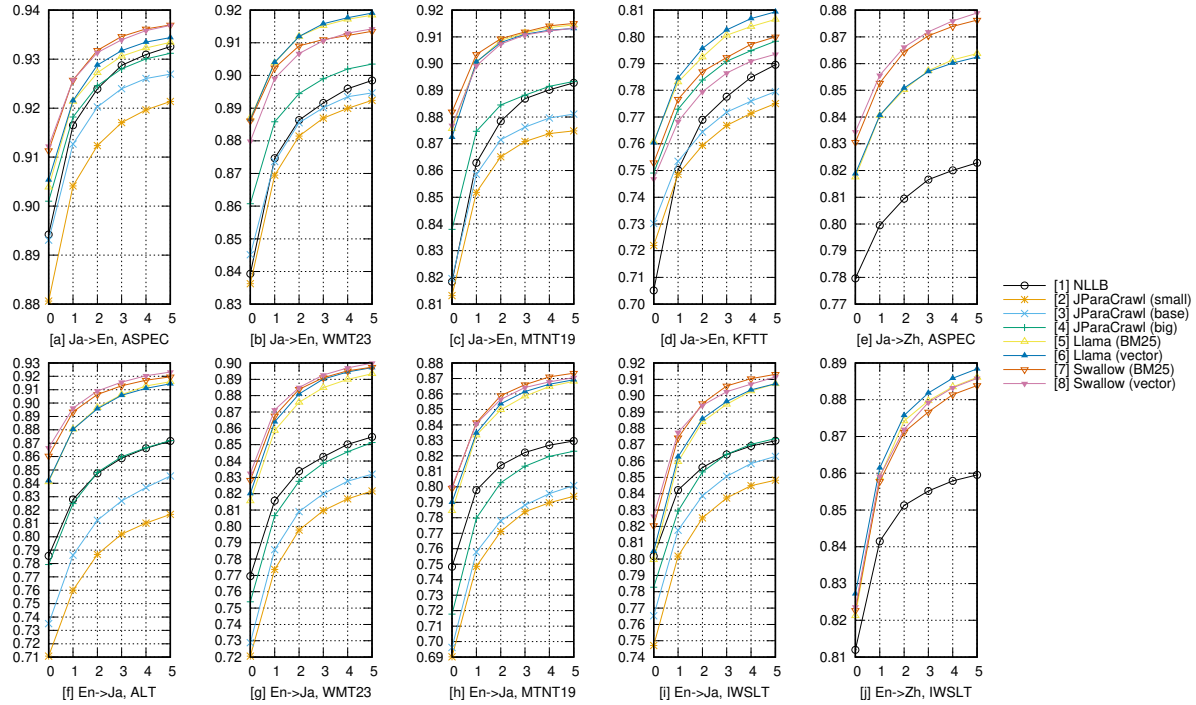


Figure 2: Translation quality estimated by XCOMET-XL without reference at each iteration of our method.

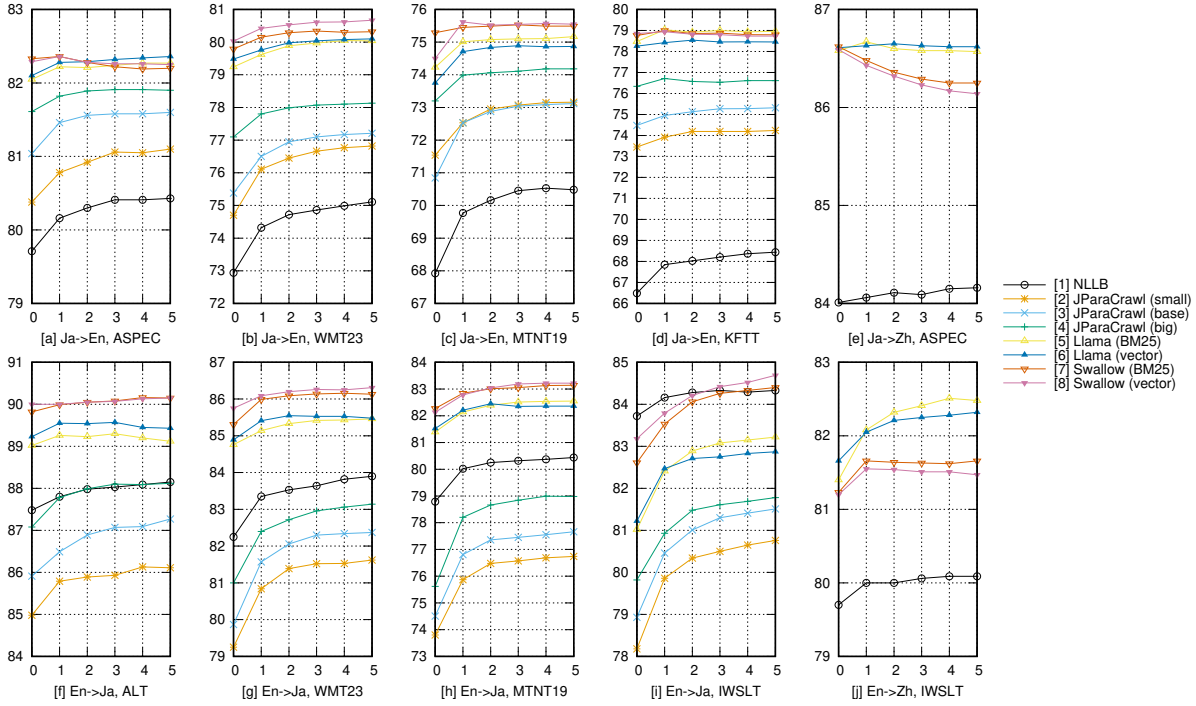


Figure 3: COMET score computed by wmt22-comet-da with a reference at each iteration of our method.

6.3 Source Text Editor

We quantified the degree of text editing performed by each method, with the translation error rate (TER) (Snover et al., 2006) at the dataset level. More specifically, we tokenized Japanese and English texts using MeCab (Kudo et al., 2004) and Moses tokenizer (Koehn et al., 2007), respectively, and computed TER using SacreBLEU (Post, 2018),[5] regarding the original and edited source texts as the reference and hypothesis, respectively.

Figure 4 shows that our method altered 9%–37% of linguistic tokens. The ratio was higher

[5]Signature: nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.5.1

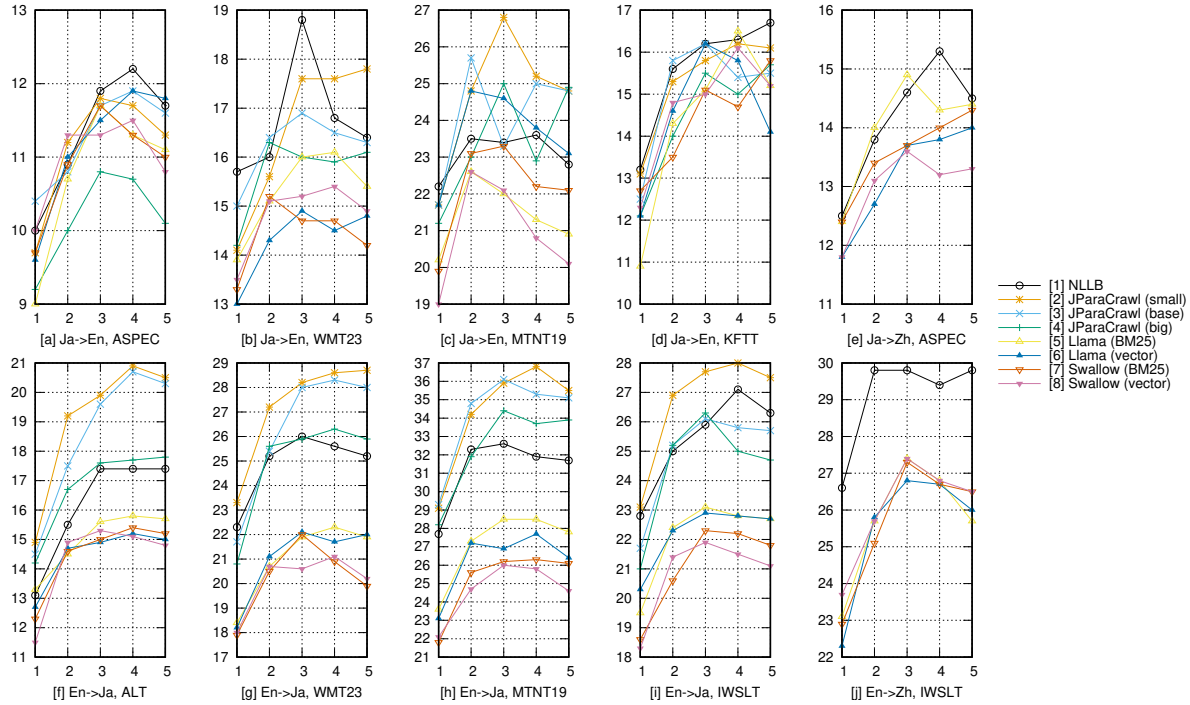


Figure 4: Translation edit rate (TER) between the original source text src_0 and each of its edited versions src_i generated by our method. Because of the search function, src_{i+1} was not necessarily obtained from src_i directly.

than those exhibited by the word-substitution method (9%–19%; Appendix F), except for the two ASPEC tasks (9%–16% vs. 17%–19%), and lower than those performed by the other two non-targeted methods (12%–76% and 19%–68% by the sequence-to-sequence and LLM-NT methods, respectively; Appendix F). We also observed that the ratio varied across the targeted MT systems.

The modest ratio of our method reflects the length distribution of the translation error spans and corresponding source text spans identified by XCOMET-XL and OTAlign. We thus consider that the LLMs properly conformed to the instruction for targeted text editing.

Note that we do not consider the ratio to be a good indicator of better translations.

7 Conclusion

As an approach to exploiting black-box MT systems, we focused on automatic and targeted source text editing. To overcome the two issues that remain in the literature, i.e., the limited ability of editing and unawareness of actual translation errors, we used LLMs, expecting that they would have sufficiently high competence to realize diverse types of edits that can improve translation quality (Miyata and Fujita, 2017, 2021) and a segment-level quality estimator as in a concurrent work (Ki and Carpuat, 2025), and implemented targeted paraphrasing (Uchimoto et al., 2006; Resnik et al., 2010) by harnessing a span-level quality estimator (error detector) and a word aligner.

Our experiments with eight MT systems and ten test datasets for four translation directions confirmed the efficacy of our method in improving translation quality, while the non-targeted counterpart (LLM-NT) also achieved a rivaling performance. Through the analyses, we identified that the impact of our method varied depending on the MT system and dataset, and that the segment-level quality estimator is the vital component that requires further improvements.

Future work includes improving each component of our method, while simplifying and speeding up the whole pipeline. Since our method focuses on source texts that lead to translation errors according to an error detector, applying LLM-NT to other error-free segments will be a straightforward way for improving translation for entire datasets. Only prompt templates are language dependent; hence, we plan to evaluate the applicability of our method to other translation tasks as well as other MT systems, including proprietary ones. We are also interested in assessing the applicability of the proposed method to other text-to-text tasks, including text summarization and text simplification.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and suggestions. A part of these research results was obtained from the commissioned research (No. 22501) by NICT.

Limitations

As observed in our experiments, the gain of the COMET score achieved by our method and competing methods depended on the target MT systems and test datasets. In addition, the COMET score may evaluate the translation quality only from limited perspectives, heavily relying on a single reference translation. Thus, our results do not guarantee the same conclusions for other MT systems, datasets, and translation directions. For instance, our method may not work well for translating from/into low-resource languages, provided that the component models of our method, i.e., quality estimator, error detector, aligner, and paraphraser, have not been trained for those languages and thereby perform less accurately.

We used COMET (wmt22-comet-da) for evaluating the translation quality, following Freitag et al. (2022); they reported that it achieved the highest correlation with human rating among reproducible automatic metrics. However, recent work, such as Agrawal et al. (2024), demonstrated that XCOMET-XL surpassed COMET. Evaluating with XCOMET-XL may lead to different conclusions. The discrepancy between reference-free and reference-based metrics observed in Figures 2 and 3 could be resolved. On the other hand, the use of the same model for search and evaluation may lead to a bias. To confirm the gain in translation quality, human evaluation is indispensable.

We made large efforts to refine the prompt templates for text editing and translation with LLMs. However, there is still room for improvement. While our prompt templates (Appendix B.1) follow the spirit of "targeted paraphrasing" (Uchimoto et al., 2006; Resnik et al., 2010), a concurrent work (Ki and Carpuat, 2025) has demonstrated that the instruction for text simplification results in better translations than instructing on the paraphrasing task. The optimized templates for an LLM may not work well for other LLMs.

LLM-based source text editing requires substantially larger computes than existing methods that do not rely on LLMs. However, we assume that the latency with $N = 5$ and eight NVIDIA Tesla V100 GPUs is acceptable for the existing translation production process (ISO/TC37, 2017), where the manual post-editing step governs latency. We expect that recent advances in smaller language models and model compression will make our approach faster and consequently more feasible.

Ethics Statements

The resulted translations had, on average, a higher quality according to the COMET score. However, translation errors should remain. Therefore, the direct use of MT outputs could mislead potential users.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502.

Yuki Arase, Han Bao, and Sho Yokoi. 2023. Unbalanced Optimal Transport for Unbalanced Word Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3966–3986.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 Evaluation Campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. *Preprint*, arXiv:2401.08281.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 878–891.

Marina Fomicheva, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao. 2021. The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results. In *Proceedings of the Second Workshop on Evaluation and Comparison of NLP Systems*, pages 165–178.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does Prompt Formatting Have Any Impact on LLM Performance? *Preprint*, arXiv:2411.10541.

ISO/TC37. 2017. ISO 18587:2017 Translation Services – Post-editing of Machine Translation Output – Requirements.

Dayeon Ki and Marine Carpuat. 2025. Automatic Input Rewriting Improves Translation with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10829–10856.

Yeun-Bae Kim and Terumasa Ehara. 1994. An Automatic Sentence Breaking and Subject Supplement Method for J/E Machine Translation. *IPSJ Journal*, 35(6):1018–1028. (in Japanese).

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 1–45.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180.

Hyuga Koretaka, Tomoyuki Kajiwara, Atsushi Fujita, and Takashi Ninomiya. 2023. Mitigating Domain Mismatch in Machine Translation via Paraphrasing. In *Proceedings of the Tenth Workshop on Asian Translation*, pages 29–40.

Taku Kudo and John Richardson. 2018. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019. Findings of the First Shared Task on Machine Translation Robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 91–102.

Xing Han Lù. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

Sneha Mehta, Bahareh Azarnoush, Boris Chen, Avneesh Saluja, Vinith Misra, Ballav Bihani, and Ritwik Kumar. 2020. Simplify-Then-Translate: Automatic Preprocessing for Black-Box Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8488–8495.

Paul Michel and Graham Neubig. 2018. MTNT: A Testbed for Machine Translation of Noisy Text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553.

Shachar Mirkin, Sriram Venkatapathy, Marc Dymetman, and Ioan Calapodescu. 2013. SORT: An Interactive Source-Rewriting Tool for Improved Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 85–90.

Rei Miyata and Atsushi Fujita. 2017. Dissecting Human Pre-Editing toward Better Use of Off-the-Shelf Machine Translation Systems. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, pages 54–59.

Rei Miyata and Atsushi Fujita. 2021. Understanding Pre-Editing for Black-Box Neural Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1539–1550.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710.

Masaaki Nagata, Makoto Morishita, Katsuki Chousa, and Norihito Yasuda. 2024. A Japanese-Chinese Parallel Corpus Using Crowdsourcing for Web Mining. *Preprint*, arXiv:2405.09017.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian Scientific Paper Excerpt Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2204–2208.

Hiroaki Nanjo, Yuji Yamamoto, and Takehiko Yoshimi. 2012. Automatic Construction of Statistical Preediting System from Parallel Corpus for Improvement of Machine Translation Quality. *IPSJ Journal*, 64(6):1644–1653. (in Japanese).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B. Bederson. 2010. Improving Translation via Targeted Paraphrasing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 127–137.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2016. Introduction of the Asian Language Treebank. In *Proceedings of the 2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, pages 1–6.

Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Satoshi Shirai, Satoru Ikehara, and Tsukasa Kawaoka. 1993. Effects of Automatic Rewriting of Source Language within a Japanese to English MT System. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 226–239.

Satoshi Shirai, Satoru Ikehara, Akio Yokoo, and Yoshifumi Ooyama. 1998. Automatic Rewriting Method for Internal Expressions in Japanese to English MT

and Its Effects. In *Proceedings of the Second International Workshop on Controlled Language Applications*, pages 62–75.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Seventh Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 Shared Task on Quality Estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 Shared Task on Quality Estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725.

Sanja Štajner and Maja Popovic. 2016. Can Text Simplification Help Machine Translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Sanja Štajner and Maja Popović. 2018. Improving Machine Translation of English Relative Clauses with Automatic Text Simplification. In *Proceedings of the First Workshop on Automatic Text Adaptation*, pages 39–48.

Yanli Sun, Sharon O'Brien, Minako O'Hagan, and Fred Hollowood. 2010. A Novel Statistical Pre-Processing Model for Rule-Based Machine Translation System. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*.

Kiyotaka Uchimoto, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2006. Automatic Detection and Semi-Automatic Revision of Non-Machine-Translatable Parts of a Sentence. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 703–708.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285.

Masaya Yamaguchi, Nobuo Inui, Yoshiyuki Kotani, and Hirohiko Nishimura. 1998. Acquisition of Automatic Pre-edition Rules from Results of Pre-edition. *IPSJ Journal*, 39(1):17–28. (in Japanese).

Takehiko Yoshimi. 2001. Improvement of Translation Quality of English Newspaper Headlines by Automatic Pre-editing. *Machine Translation*, 16:233–250.

A Public Resources Used

Table 4 lists the links to the resources, including the datasets, pre-trained model checkpoints, and tools, used in our experiments (Sections 4–6).

B Prompt Templates

B.1 For Targeted Source Text Editing

Figure 5 presents the prompt templates used for our targeted source text editing. To perform this task, we used an LLM that is mainly trained on the language of interest, i.e., Llama-Swallow for Japanese and Llama for English, and provided prompts in the same language. Pairs of double brackets ("{{" and "}}") indicate placeholders. Given a source text to be edited, a tailored prompt is automatically instantiated by filling these placeholders.

B.2 For Translation

Figure 6 presents the prompt templates used for MT. In the same manner as for source text editing, we provided prompts in the language for which the LLM was recently and mainly trained on, i.e., Japanese for Llama-Swallow and English for Llama. The role of double brackets is the same as for source text editing.

C Preliminary Experiments

As described in §4, we explored a better way of determining the source text span to be edited and the LLM used for source text editing, using the sampled set of 100 parallel sentences and the JParaCrawl Japanese-to-English NMT model (big).

C.1 Source Text Span Detection Methods

To confirm the feasibility and impact of determining source text spans to be edited, we compared the following three methods.

Random: This method first specifies the beginning of the source text span randomly, and then the end from the remainder of the source text.

Datasets
Asian Language Treebank (ALT), https://huggingface.co/datasets/mutiyama/alt
ASPEC, https://jipsti.jst.go.jp/aspec/
IWSLT 2017, https://huggingface.co/datasets/IWSLT/iwslt2017
Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles, https://alaginrc.nict.go.jp/WikiCorpus/, 2.01
JParaCrawl, https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/, v3.0
Kyoto Free Translation Task (KFTT), https://www.phontron.com/kftt/, 1.4
MTNT, https://github.com/pmichel31415/mtnt/, v1.1
MTNT19, https://pmichel31415.github.io/mtnt/, MTNT2019.tar.gz
WMT22 Test sets, https://github.com/wmt-conference/wmt22-news-systems, v1.1
WMT23 Test sets, https://github.com/wmt-conference/wmt23-news-systems, v0.1
Pre-trained model checkpoints
FastText word embeddings, https://fasttext.cc/docs/en/crawl-vectors.html
Google BERT large, https://huggingface.co/google-bert/bert-large-cased-whole-word-masking
InfoXLM-Base, https://huggingface.co/microsoft/infoxlm-base
JParaCrawl NMT Models, https://www.kecl.ntt.co.jp/icl/lirg/jparacrawl/, based on v3.0
LaBSE, https://huggingface.co/sentence-transformers/LaBSE
Llama-3.1-70B-Instruct, https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
Llama-3.1-8B-Instruct, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Llama-3.1-Swallow-70B-Instruct-v0.1, https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-70B-Instruct-v0.1
Llama-3.1-Swallow-8B-Instruct-v0.1, https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.1
M2M100-418M, https://huggingface.co/facebook/m2m100_418M
NLLB-200-3.3B, https://huggingface.co/facebook/nllb-200-3.3B
Tohoku-NLP BERT base Japanese, https://huggingface.co/tohoku-nlp/bert-base-japanese-whole-word-masking
wmt22-comet-da, https://huggingface.co/Unbabel/wmt22-comet-da
XCOMET-XL, https://huggingface.co/Unbabel/XCOMET-XL
Tools
bm25s, https://github.com/xhluca/bm25s, 0.2.6
COMET, https://github.com/Unbabel/COMET, 2.2.5
Fairseq, https://github.com/facebookresearch/fairseq, v0.12.2
Faiss, https://github.com/facebookresearch/faiss, v1.7.2
Language Model Evaluation Harness, https://github.com/EleutherAI/lm-evaluation-harness, v0.4.3
MeCab, https://taku910.github.io/mecab/, 0.996
Moses tokenizer, https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl, RELEASE-4.0
OTAlign, https://github.com/yukiar/OTAlign, 79fefaa
SacreBLEU, https://github.com/mjpost/sacrebleu, v2.5.1
SentencePiece, https://github.com/google/sentencepiece, v0.2.0

Table 4: Public resources used in our experiments.

Direct: The error detector, XCOMET-XL, annotates only erroneous text spans in the hypothesis. This method examines whether it can directly annotate the source text spans by swapping the source text and hypothesis.

Propagation: A combination of XCOMET-XL for annotating erroneous text spans in the hypothesis and a word aligner, OTAlign, to propagate those spans to the source text.

Figure 7 shows the results with Llama-Swallow for Japanese source text editing (Appendix C.2). The "Propagation" method achieved the highest COMET score with any value of N up to 5, even though it should have involved prediction errors of both the error detector and word aligner. Interestingly, with the "Random" text spans, our method improved the COMET score up to 1.0 point. Even though the "Direct" assessment of the source text led to a higher COMET score than "Random," it lagged behind "Propagation."

We thus chose the "Propagation" method in our experiments in §5, whose prediction could become even more accurate if the two components are improved. In addition, in the literature on translation quality estimation, researchers have attempted to determine source text spans corresponding to translation errors (Specia et al., 2020, 2021; Fomicheva et al., 2021). Although this line of research is out of scope in the recent series of shared tasks, we believe it is worth considering, in particular for promoting source text editing.

C.2 LLMs for Text Editing

A number of LLMs are publicly available, but the extent to which they perform the source text editing task is unknown. We considered that instruction tuning is necessary. Hence, we compared the following four LLMs that differ in the existence of language-specific adaptation and model size.

Llama-Swallow-70B: The Llama model with 70B parameters, continually pre-trained on

```
日本語文の言い換え文を出力してください。
手順は次の通りです。
1. 日本語文の¶¶で囲まれた言い換え対象表現について、同じ意味を持つ異なる表現の言い換え候補を1個から5個挙げてください。
2. 手順1で挙げた言い換え候補の中から、日本語文の¶¶で囲まれた箇所を言い換えるのに、最も適切な候補を1つ選んでください。
3. 日本語文の¶¶で囲まれた箇所を、手順2で選ばれた最適な言い換え候補を使用して置換してください。この際、以下の基準を満たすように文脈に合わせて適切に調整してください。
・文脈に応じて、適切な動詞の活用形や助詞を使うこと。
・元の文の意味を正確に伝えること。
・文法的に正しい構文を持つこと。

入力例:
日本語文: {{paraphrase["example_original"]}}
言い換え対象表現: {{paraphrase["example_original_span"]}}
出力例:
{"言い換え文":"{{paraphrase["example_paraphrase"]}}"}

日本語文: {{paraphrase["annotated_src"]}}
言い換え対象表現: {{paraphrase["propagate_error_span"]}}
```

(a) Prompt template used for Llama-Swallow to edit Japanese source text.

```
Please output a paraphrased sentence for a given English sentence.
The procedure is as follows:
1. For the target expression for paraphrasing marked with ¶¶ in the English sentence, provide 1 to 5 paraphrase candidates with the same meaning and different expressions.
2. Select one among the paraphrase candidates generated in step 1 that is most appropriate for paraphrasing the part marked with ¶¶ in the English sentence.
3. Replace the part marked with ¶¶ in the English sentence with the paraphrase candidate selected in step 2. At the same time, please perform necessary adjustment to make it fit the context while meeting the following criteria.
・Use appropriate conjugation form of words and particles according to the context.
・Convey the original meaning of the sentence accurately.
・Maintain the grammatically correct structure of the sentence.

Input example:
English sentence: {{paraphrase["example_original"]}}
Target expression for paraphrasing: {{paraphrase["example_original_span"]}}
Output example:
{"paraphrased_sentence":"{{paraphrase["example_paraphrase"]}}"}

English sentence: {{paraphrase["annotated_src"]}}
Target expression for paraphrasing: {{paraphrase["propagate_error_span"]}}
```

(b) Prompt template used for Llama to edit English source text.

Placeholder	Content to be filled
paraphrase["example_original"]	Source text of the retrieved example
paraphrase["example_original_span"]	Targeted span in the above text
paraphrase["example_paraphrase"]	Paraphrased text of the retrieved example
paraphrase["annotated_src"]	Source text to be edited
paraphrase["propagated_error_span"]	Targeted span in the above text to be edited

(c) Placeholders in the prompt templates for text editing.

Figure 5: Prompt templates used for source text editing.

a massive text data in Japanese (Fujii et al., 2024).

Llama-Swallow-8B: A smaller model obtained in the same manner as above.

Llama-70B: The 70B parameter model not specially adapted to Japanese (Grattafiori et al., 2024).

Llama-8B: A smaller model obtained in the same manner as above.

We also evaluated manual source text editing performed by the first author, which cannot be the upper bound, but is a good reference. Note that other components, including source text span detection method and prompt templates, were the same as our final method.

Figure 8 demonstrates that the four LLMs and human editor had a clear order in the COMET score

```
{{source_language}}を{{target_language}}に翻訳してください。

入力例:
英語文: {{translation["example_src"]}}
出力例:
{"翻訳文":"{{translation["example_tgt"]}}"}

英語文: {{translation["src"]}}
```

(a) Prompt template used for translation with Llama-Swallow.

```
Please translate the {{source_language}} sentence into {{target_language}}.

Input example:
Japanese sentence: {{translation["example_src"]}}
Output example:
{"translation":"{{translation["example_tgt"]}}"}

{{source_language}} sentence: {{translation["src"]}}
```

(b) Prompt template used for translation with Llama.

Placeholder	Content to be filled
source_language	Source language (e.g., "Japanese")
target_language	Target language (e.g., "English")
translation["example_src"]	Source text of the retrieved translation example
translation["example_tgt"]	Target text of the retrieved translation example
translation["src"]	Source text to be translated

(c) Placeholders in the prompt templates for translation.

Figure 6: Prompt templates used for translation.

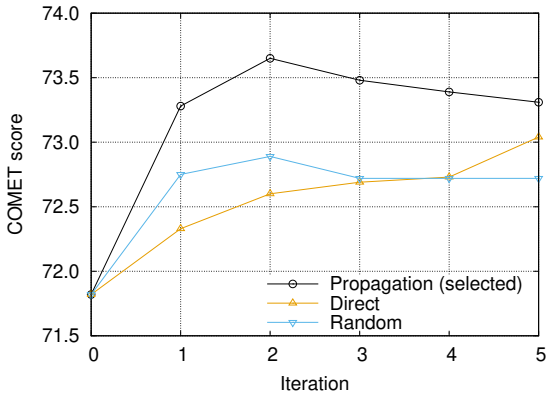


Figure 7: COMET scores achieved by different source text span detection methods.

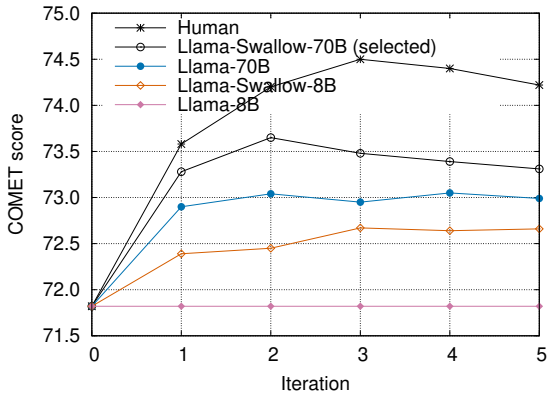


Figure 8: COMET scores achieved by different source text editors.

with any value of N up to 5. When we compared LLMs with the same sizes, Llama-Swallow outperformed Llama, which confirms the benefit of adaptation to the language of interest. We also found that smaller LLMs were not as good as their larger counterpart. Although some coincidences could exist, the non-adapted small LLM, i.e., Llama-8B, did not improve the COMET score at all. Following the results, we used Llama-Swallow-70B for editing source texts in Japanese, and analogously Llama-70B for English (§5.1.1).

At the time of this preliminary experiment, we required a sufficiently large model. However, recent advances in smaller language models will lead to a better balance of cost and benefit.

D Erroneous Test Subsets

Our method attempts to avoid translation errors. Thus, we mainly evaluated and analyzed our method focusing on the erroneous test subsets determined using XCOMET-XL (§5.1.3).

Table 5 summarizes the number of lines containing at least one "critical," "major," or "minor" error when translating with each MT system.

MT System	Ja→En				Ja→Zh	En→Ja				En→Zh
	ASPEC [a]	WMT23 [b]	MTNT19 [c]	KFTT [d]	ASPEC [e]	ALT [f]	WMT23 [g]	MTNT19 [h]	IWSLT [i]	IWSLT [j]
NLLB [1]	1,233	1,332	766	961	2,026	823	1,638	1,077	1,097	1,207
JParaCrawl (small) [2]	1,314	1,544	894	1,069	-	900	1,817	1,225	1,215	-
JParaCrawl (base) [3]	1,312	1,514	888	1,065	-	879	1,802	1,214	1,211	-
JParaCrawl (big) [4]	1,224	1,432	862	1,077	-	844	1,755	1,225	1,163	-
Llama (BM25) [5]	1,261	1,321	795	1,037	2,003	767	1,630	1,126	1,114	1,264
Llama (vector) [6]	1,253	1,331	789	1,040	1,979	748	1,612	1,110	1,101	1,272
Llama-Swallow (BM25) [7]	1,141	1,262	732	1,029	1,940	729	1,538	1,072	1,064	1,295
Llama-Swallow (vector) [8]	1,117	1,250	733	1,035	1,914	716	1,525	1,066	1,046	1,306
All	1,812	1,992	1,110	1,160	2,107	1,018	2,074	1,392	1,452	1,459

Table 5: Number of lines containing "critical," "major," or "minor" errors in the baseline translation detected by XCOMET-XL.

MT System	Ja→En				Ja→Zh	En→Ja				En→Zh
	ASPEC [a]	WMT23 [b]	MTNT19 [c]	KFTT [d]	ASPEC [e]	ALT [f]	WMT23 [g]	MTNT19 [h]	IWSLT [i]	IWSLT [j]
NLLB [1]	0.492	0.528	0.609	0.614	0.329	0.389	0.357	0.392	0.357	0.392
JParaCrawl (small) [2]	0.476	0.489	0.591	0.682	-	0.427	0.349	0.404	0.292	-
JParaCrawl (base) [3]	0.442	0.527	0.553	0.634	-	0.354	0.346	0.353	0.296	-
JParaCrawl (big) [4]	0.518	0.527	0.523	0.741	-	0.429	0.331	0.425	0.296	-
Llama (BM25) [5]	0.582	0.545	0.481	0.663	0.386	0.581	0.400	0.453	0.373	0.448
Llama (vector) [6]	0.530	0.576	0.578	0.637	0.368	0.617	0.441	0.440	0.466	0.436
Llama-Swallow (BM25) [7]	0.762	0.617	0.682	0.753	0.534	0.356	0.428	0.381	0.465	0.493
Llama-Swallow (vector) [8]	0.703	0.638	0.646	0.682	0.519	0.428	0.409	0.395	0.433	0.418

Table 6: Pearson product-moment correlation coefficients r between the COMET gain (wmt22-comet-da, with reference) and QE score gain (XCOMET-XL, without reference) both achieved by the proposed method. See Table 5 for the number of segments in each configuration.

E Correlation between the Estimated Quality and COMET Score

Table 6 summarizes the segment-level correlation coefficients between the gain of the COMET score and the gain of the estimated quality. At a glance, a moderate positive correlation existed for most configurations, and one might consider that this indicates that the estimated quality was beneficial to the search. However, as observed in §6.2, the estimated score does not always help the system make the correct decision.

F Source Text Edit Rate

Figures 9, 10, and 11 show the degree of text editing performed by the word-substitution, sequence-to-sequence, and LLM-NT methods, respectively, measured by TER computed in the same manner as for our method (§6.3). Compared with our method, the word-substitution method led to a lower TER, since it substituted only one word. The other two non-targeted methods indiscriminately affected the entire texts, which led to a substantially larger TER depending mainly on the dataset: the sequence-to-sequence method for Japanese and the LLM-NT method for English. LLM-NT demonstrated a saturation, indicating that it properly followed the instruction for retaining semantics and grammaticality.

G Details of the Sequence-to-Sequence Paraphrasing Models

Only the monolingual sequence-to-sequence pre-editing models were trained by us for our experiments. Our procedure below follows that in a previous study (Koretaka et al., 2023).

First, we generated synthetic monolingual parallel data from the bilingual parallel corpora used for retrieving translation demonstrations (§5.1.2). We randomly sampled text pairs from each corpus, aligning their sizes with the minimum corpus of 4.6M, and translated their non-targeted side into the language on the other side, using M2M100-418M (Fan et al., 2021) and beam search with a beam size of 12. The synthetic monolingual parallel data, 9.2M text pairs for each of Japanese and English, were composed of the pairs of the resulted translation (Ja'←{En,Zh} and En'←{Ja,Zh}) and the corresponding reference translation in the bilingual parallel corpus (Ja–{En,Zh} and En–{Ja,Zh}).

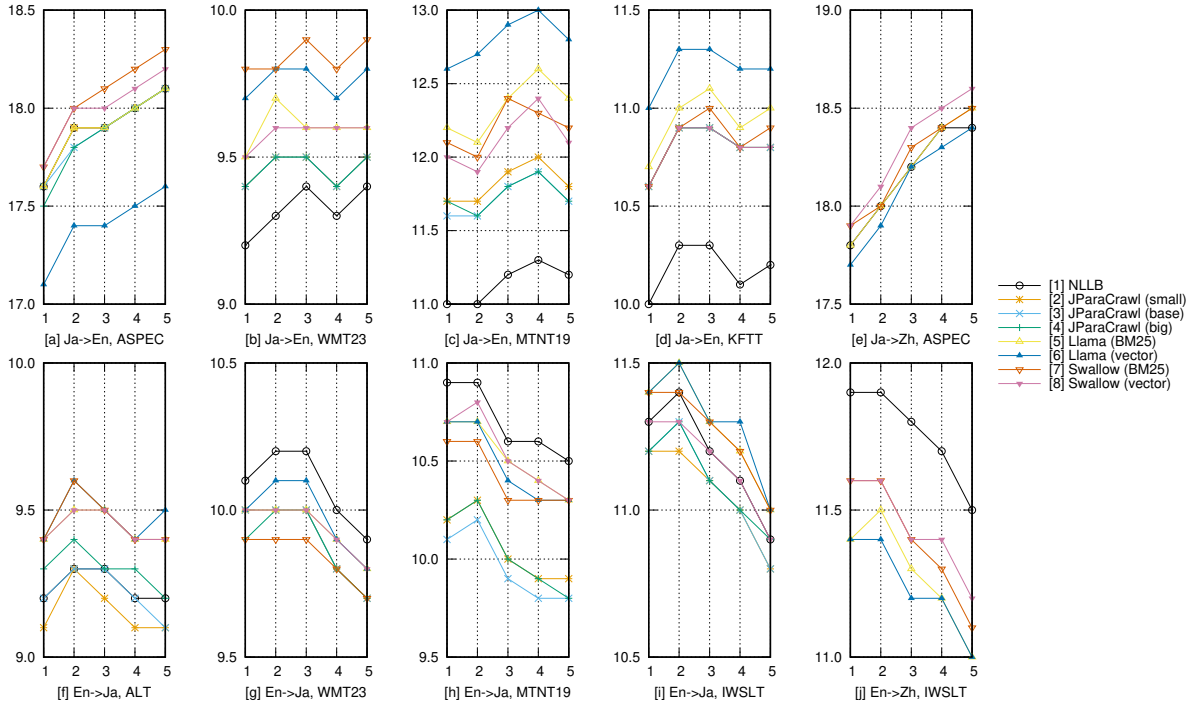


Figure 9: Translation edit rate (TER) between the original source text src_0 and each of its edited versions src_i generated by the Word-Sub method (the five-best outputs).

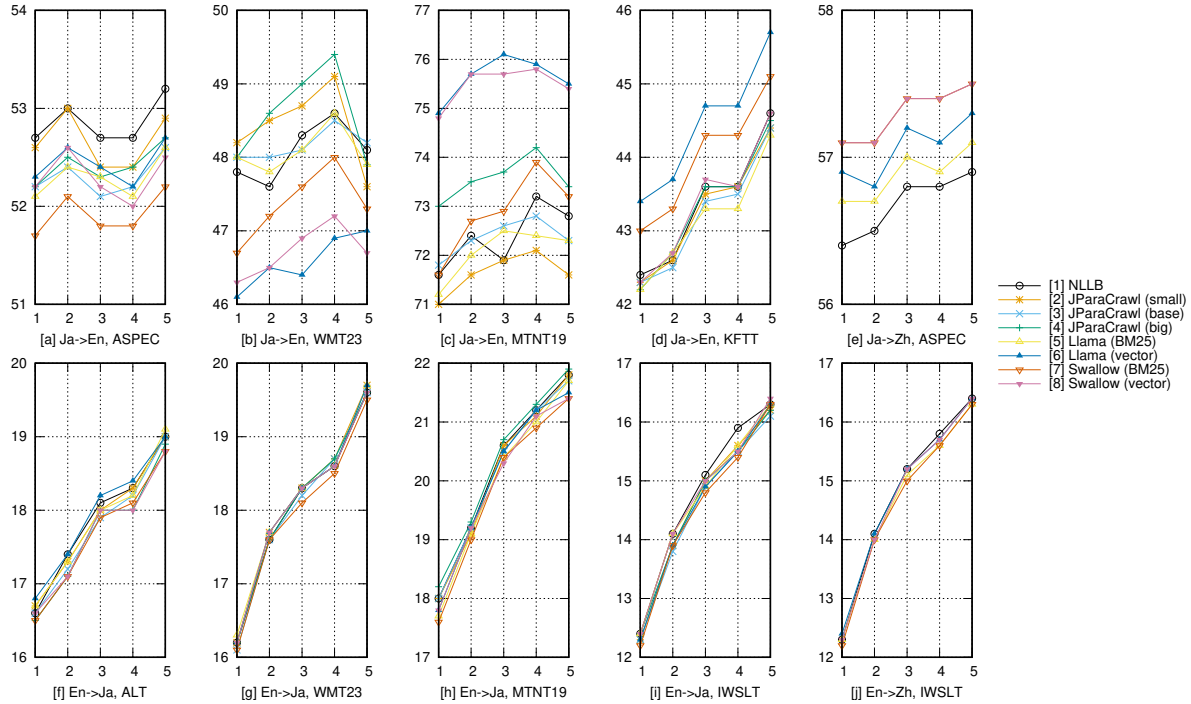


Figure 10: Translation edit rate (TER) between the original source text src_0 and each of its edited versions src_i generated by the Seq2seq-B method (the five-best outputs).

We then trained a Transformer Base model (Vaswani et al., 2017) for each of Japanese and English, regarding the synthetic side as the source, and using a joint vocabulary of 32k sub-words determined using SentencePiece (Kudo and Richardson, 2018) and Fairseq (Ott et al., 2019). We set the training hyper-parameters to the same values as Morishita et al. (2022), except for the number of updates of 60k and the lack of checkpoint averaging.

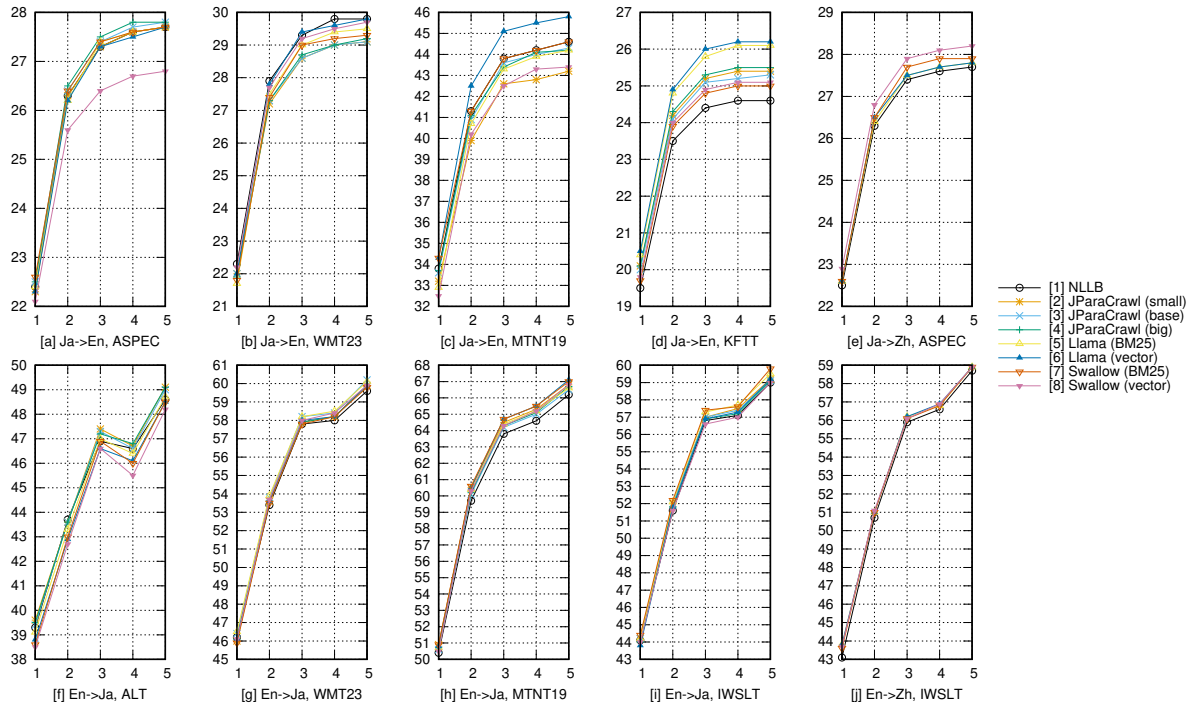


Figure 11: Translation edit rate (TER) between the original source text src_0 and each of its edited versions src_i generated by the LLM-NT method. Because of the iterative nature, src_{i+1} was always obtained from src_i directly.

Self-Retrieval from Distant Contexts for Document-Level Machine Translation

Ziqian Peng[1,2] and **Rachel Bawden**[2] and **François Yvon**[1]
[1]Sorbonne Université & CNRS, ISIR, Paris, France
[2]Inria, Paris, France
{ziqian.peng,francois.yvon}@isir.upmc.fr rachel.bawden@inria.fr

Abstract

Document-level machine translation is a challenging task, as it requires modeling both short-range and long-range dependencies to maintain the coherence and cohesion of the generated translation. However, these dependencies are sparse, and most context-augmented translation systems resort to two equally unsatisfactory options: either to include maximally long contexts, hoping that the useful dependencies are not lost in the noise; or to use limited local contexts, at the risk of missing relevant information. In this work, we study a self-retrieval-augmented machine translation framework (SELF-RAMT), aimed at informing translation decisions with informative local and global contexts dynamically extracted from the source and target texts. We examine the effectiveness of this method using three large language models, considering three criteria for context selection. We carry out experiments on TED talks as well as parallel scientific articles, considering three translation directions. Our results show that integrating distant contexts with SELF-RAMT improves translation quality as measured by reference-based scores and consistency metrics.

1 Introduction

Document-level machine translation (DLMT) is a challenging task, as it requires modeling both short-range and long-range dependencies to maintain the coherence and cohesion of the generated translation. Inter-sentential contexts are indispensable for the handling of phenomena such as co-reference, lexical consistency, textual coherence and cohesiveness, which continue to be challenging for long document translation (Bawden et al., 2018; Maruf et al., 2019; Voita et al., 2019b; Fernandes et al., 2023). Numerous approaches, reviewed in (Maruf et al., 2021; Castilho and Knowles, 2024), have been proposed to integrate these contexts. They include segment concatenation (Tiedemann and Scherrer, 2017; Bawden et al., 2018; Sun et al., 2022), architecture adaptation (Miculicich et al., 2018; Yang et al., 2019; Ma et al., 2020), training strategy optimization (Lupo et al., 2022b; Li et al., 2023; Wu et al., 2024), and multi-pass refinement (Voita et al., 2019a; Yu et al., 2020; Koneru et al., 2024). Past work also shows that various sources of contextual information contribute differently to translation quality; the local source and target context is the main resource for handling anaphoric references and word-sense disambiguation information (Bawden et al., 2018; Gete et al., 2022), whereas the global context, especially on the target side, holds information likely to improve coherence and cohesiveness of the full translated document (Pal et al., 2024).

Recent generative models such as Llama3 (Grattafiori et al., 2024) and GPT4 (OpenAI et al., 2023) can process inputs up to hundreds of thousands of tokens, creating new possibilities for the inclusion of the whole source text, as well as already translated target segments, in the translation context. It however remains an open question whether such architectures, relying on the self-attention-mechanism (Vaswani et al., 2017), are effectively able to identify relevant long-range dependencies and actually improve DLMT (Wang et al., 2024). This is because inter-sentential dependencies can be sparsely distributed within a document, whereas self-attention generates dense patterns spreading out over the entire past text (Tay et al., 2023; Liu et al., 2024). Therefore, most approaches to DLMT still consider a limited context window size, usually up to 1024 tokens or a fixed number of sentences.

In order to capture long-distance dependencies without requiring the attention mechanism to handle overly long contexts, we propose self-retrieval augmentation for machine translation (SELF-RAMT), aiming to take into account both local and global dependencies, regardless of the

Proceedings of the Tenth Conference on Machine Translation, pages 220–240
November 8-9, 2025

document length. In our approach, inter-sentential dependencies are precomputed for the full source document to identify past relevant segments for each translation unit. As soon as they are translated, these segments and their translation become available to inform the subsequent translation choices. In our implementation, which uses large language models (LLMs), such dynamic contexts are taken into account through in-context learning (ICL). Two scenarios are considered: (a) one where in-context examples correspond to correct translations, as in online learning, where the input sentences, once incrementally post-edited by a human translator, become available to revise the model (Álvaro Peris and Casacuberta, 2019),[1] and (b) a fully automatic setup, with imperfect in-context translations, requiring no human intervention. In this context, our main research questions are as follows: (i) how to best identify and retrieve useful context segments, (ii) what improvement to MT quality do these retrieved contexts bring, and (iii) to what extent distant (as opposed to local) contexts actually enhance translation scores. We compare three criteria for context selection (cosine similarity with respect to LaBSE embeddings (COS), Best Match 25 (BM25) and point-wise mutual information (PMI)) and carry out experiments on three LLMs in three language directions (English to German (EN–DE), French (EN–FR), and Chinese (EN–ZH)), analyzing the impact of contexts, especially distant ones. Experimental data includes both TED talks (Cettolo et al., 2012) and a new dedicated parallel test set, MERSENNE, consisting of scientific articles for the EN–FR direction. Scientific articles offer an interesting use case to study term consistency in long document translation. Our investigation reveals that distant contexts retrieved with PMI provide valuable information that increases translation metric scores as well as term consistency. We make our code and data available.[2]

2 Related Work

Document-level MT DLMT research broadly falls into two categories: *Doc2Sent*, which involves translating each sentence individually using intra-document source and/or target context to aid translation, and *Doc2Doc*, which involves translating multiple sentences at once (Popescu-Belis, 2019; Maruf et al., 2021; Castilho and Knowles, 2024). *Doc2Doc* approaches represent a simple strategy for effective context integration, maintaining better consistency and coherence than *Doc2Sent* methods (Li et al., 2020; Sun et al., 2022). However, *Doc2Doc* methods struggle to process very long sequences of sentences, as relevant information is sparse in global contexts (Lupo et al., 2022a; Wang et al., 2023), which can lead to the degradation of translation quality or omitted sentences (Zhuocheng et al., 2023; Li et al., 2023; Peng et al., 2025). To address this problem, several approaches have been explored, including context-aware attention (Maruf et al., 2019; Zheng et al., 2021; Yang et al., 2023), which selects important contexts according to the attention distribution, and dynamic context selection (Kang et al., 2020), which applies a reward model to identify varying numbers of useful context sentences within the local context, constrained by the complexity of reinforced training.

LLM-based DLMT Various LLM-based methods have been proposed for DLMT. Several works explore zero-shot prompting and ICL for *Doc2Doc* MT (Hendy et al., 2023; Karpinska and Iyyer, 2023) and study the best way to train LLMs for DLMT (Xu et al., 2024; Li et al., 2024; Guo et al., 2024; Alves et al., 2024), illustrating the importance of high-quality in-context demonstrations and fine-tuning parallel corpora. LLMs have also been used as post-editors, either through fine-tuning (Koneru et al., 2024; Li et al., 2025; Dong et al., 2025) or prompting for iterative translation refinement (Briakou et al., 2024; Wang et al., 2025a).

When it comes to integrating context, LLMs offer greater flexibility than traditional neural machine translation models. Various multi-aspect prompting techniques have been proposed to enhance the input by incorporating or automatically summarizing relevant information from the context. For instance, DELTA (Wang et al., 2025b) builds a dynamic context for each source sentence, heavily relying on LLM components to extract and assemble relevant information from the available source and target texts (including proper nouns, bilingual summaries of local contexts and relevant past sentences within a predefined context window). However, DELTA is computationally costly (a lot more so than SELF-RAMT), due to the multiple steps required for dynamic context extraction. SENT2SENT++ (Guo et al., 2025) incorporates two types of contexts: a static part, consisting of an au-

[1] We refer to this scenario as *online in-context learning*.
[2] https://anr-matos.github.io/resources.

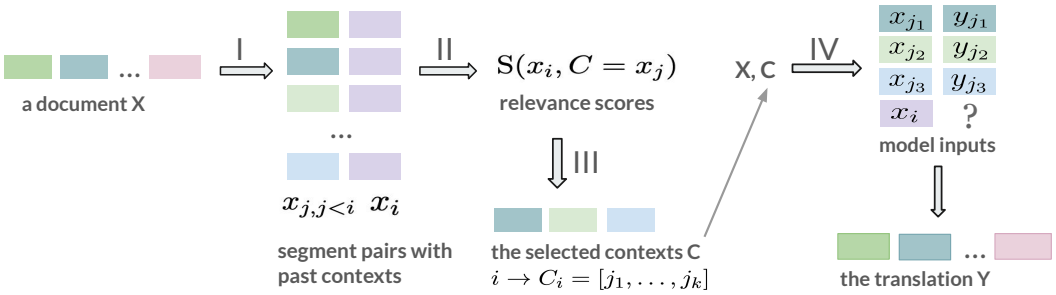


Figure 1: SELF-RAMT framework. It consists of four steps: I. Define the search domain (e.g. the past contexts); II. Compute contextual relevance scores for segments in the search domain; III. Retrieve relevant contexts according to the ranked scores and IV. Integrate the selected contexts to the inputs then generate the translations.

tomatically generated bilingual summary of the full source document, and a dynamic part, composed of the previous source and target sentences. However, the use of static, automatic summaries (a) has the effect of potentially changing the words of the context, which can be detrimental to lexical consistency and (b) does not ensure that the most relevant information is accessible for each source sentence, especially as documents increase in length.

Retrieval-augmented MT Choosing which context to be included in MT can be seen as a type of retrieval-augmented MT. In past works, retrieval-augmented MT systems have mostly been designed to mimic the use of *translation memories* by translators, which has a long history in MT (Kay, 1997). Recent implementations of this idea for neural models encode the target side of relevant example(s) together with the source sentence in an extended translation context (Gu et al., 2018; Bulte and Tezcan, 2019; Xia et al., 2019; Xu et al., 2020; He et al., 2021; Cheng et al., 2022). Variants, relying on both the source and target sides of the retrieved example(s) are proposed by Pham et al. (2020) and Reheman et al. (2023).

LLM-based MT systems seamlessly accommodate examples through in-context learning, where examples of the translation task (the source and target sides of parallel samples) (Radford et al., 2019) are inserted into the prompt. The optimal selection of in-context examples has also been the focus of recent research (Moslem et al., 2023; Vilar et al., 2023; Zhang et al., 2023; Bawden and Yvon, 2023; Agrawal et al., 2023; Cui et al., 2024; Zebaze et al., 2025), also analyzed by Zaranis et al. (2024) and Bouthors et al. (2024).

A key difference with SELF-RAMT is that these methods retrieve examples from external resources, instead of the input sequence, with the aim to find similar examples that can be easily edited. Several retrieval-augmented architectures have also been proposed, e.g., by Rubin and Berant (2024), to retrieve relevant contextual information from very long input documents. These approaches have been evaluated in language modeling tasks, but, to the best of our knowledge, have not yet been applied to MT.

3 Augmenting MT with Self-retrieval

3.1 A Self-retrieval Framework for MT

As illustrated in Figure 1, SELF-RAMT involves translating each segment of an input document X with relevant contexts retrieved within X. It consists of four steps:

I. Defining the Search Domain We consider a Doc2Sent scenario, translating sentences $\mathbf{x}_i$ in a document $X = \langle \mathbf{x}_1 \ldots \mathbf{x}_T \rangle$ using previous context sentences $\{\mathbf{x}_j, j < i\}$ in X.

II. Contextual Relevance Scores We compute contextual relevance scores $\mathrm{S}(\mathbf{x}_i, \mathbf{x}_j)$ of candidate segments $\mathbf{x}_j$ for each $\mathbf{x}_i$, with the aim of improving the consistency and coherence of the resulting translations. Details are in Section 3.2.

III. Context Retrieval For each $\mathbf{x}_i$, $\mathbf{x}_j$ is selected as a contextual segment if $\mathrm{S}(\mathbf{x}_i, \mathbf{x}_j)$ is among the top K relevance scores. Additionally, $\mathbf{x}_j$ is disregarded if $\mathrm{S}(\mathbf{x}_i, \mathbf{x}_j) \leq \tau$, where τ represents the minimum value such that $\mathbf{x}_j$ is relevant to $\mathbf{x}_i$ for score S (see Section 3.2). The resulting list of selected sentences, which we refer to as C_i, constitutes a dynamic context containing up to K sentences.

IV. Context-aware Translation Contextual sentences selected in step III are included as few-shot demonstrations in the LLM prompt in the order

in which they appear in the original text. We use specific prompts for each LLM. Details about the prompt selection and the decoding process are given in Appendix B.

3.2 Context Selection Criteria

Multiple criteria can be used to identify the relevant contextual sentences (reviewed in (Bouthors et al., 2024) for retrieval-based MT). In our approach, contextual relevance is assessed based on source side similarity between segments, which enables us to pre-compute the relevant contexts for all $\mathbf{x}_i$ prior to translation. Our hypothesis is that if $\mathbf{x}_j, j < i$ is sufficiently similar to $\mathbf{x}_i$, then $(\mathbf{x}_j, \mathbf{y}_j)$ will contain useful information when generating $\mathbf{y}_i$. This enables us to vary the retrieval score while keeping the translation infrastructure unchanged. It is therefore simpler than the proposal of Wang et al. (2025b), where contexts are dynamically updated during the generation process. In our experiments, we consider three contextual relevance scores:

COS We compute the cosine similarity between the sentence embeddings using LaBSE (Feng et al., 2022). Only positive cosine similarities are taken into account (i.e. $\tau = 0$).

BM25 We adapt BM25L (Lv and Zhai, 2011), a variant with length normalisation[3] of the Best Match 25 (BM25) relevance score (Robertson and Zaragoza, 2009), which is a go-to method for retrieving lexically relevant segments in large data stores. Implementation details are in Appendix B.2.

PMI To better reflect contextual relevance, we also consider an alternative inspired by the Pointwise Cross-Mutual Information (P-CXMI) (Fernandes et al., 2021, 2023) and the likelihood difference (Shi et al., 2024; Pombal et al., 2024). We identify relevant contexts based on the point-wise mutual information (PMI) between $\mathbf{x}_i$ and $\mathbf{x}_j$, defined as:

$$\mathrm{PMI}(\mathbf{x}_i, \mathbf{x}_j) = -\frac{1}{l_i} \sum_{t=1}^{l_i} \log \frac{P_C(x_{i,t}|x_{i,<t})}{P_C(x_{i,t}|x_{i,<t}, \mathbf{x}_j)},$$

where l_i is the length of $\mathbf{x}_i$. In other words, $\mathrm{PMI}(\mathbf{x}_i, \mathbf{x}_j)$ measures how much the knowledge of $\mathbf{x}_j$ reduces the uncertainty about $\mathbf{x}_i$ for some autoregressive language model P_C. We disregard $\mathbf{x}_j$ if $\mathrm{PMI}(\mathbf{x}_i, \mathbf{x}_j) \leq \tau$ with $\tau = 0$.

[3]Our code uses the Python implementation of Lù (2024).

Baseline We compare these relevance scores to four baselines: (i) Zero-shot, vanilla sentence-level MT, reflecting the basic non-contextual MT ability of LLMs, (ii) Past-K, where the local context is composed of K previous sentences, (iii) Random-K, where we randomly select K past sentences, and (iv) Indep-K (K independent examples generated by an LLM, the same for all sentences). More details are given in Appendix B.

4 Experimental Settings

4.1 Datasets

Our experiments rely on two test sets described below.

IWSLT Following Wang et al. (2025b) and Guo et al. (2025), we take the test sets of IWSLT2017[4] (Cettolo et al., 2012) as our test sets, for three translation directions: English to German (DE), French (FR) and Chinese (ZH), with respectively 10, 12 and 12 TED talks.

MERSENNE Due to the scarcity of long document-level data, we curated a set of 23 published scientific articles and their translations for the EN–FR language pair.[5] These articles are segmented into sentences and aligned into parallel articles. We refer to this test set as MERSENNE. More details on data preparation are given in Appendix A.

Statistics of test sets Table 1 reports the number of full documents and the number of sentences in our test sets. It also includes the average, minimum, and maximum length of sentences in LLAMA tokens, for the source and target languages. The average number of sentences is 119 for TED talks from IWSLT, and 192 for articles from MERSENNE.

4.2 Models and Inference Settings

We evaluate our framework with three open-weight medium-size multilingual LLMs using ICL: Llama3.1-8B-Instruct[6] (LLAMA) (Grattafiori et al., 2024), EuroLLM-9B-Instruct[7] (EUROLLM) (Martins et al., 2024, 2025), and Qwen2.5-7B-Instruct[8] (QWEN) (Qwen et al., 2025). These models do

[4]https://wit3.fbk.eu/2017-01-d

[5]https://www.centre-mersenne.org/

[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7]https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[8]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

	IWSLT			MERSENNE
	en-de	en-fr	en-zh	en-fr
#doc	10	12	12	23
#sent	1138	1455	1459	4417
mean	20/25	21/26	20/24	36/53
min	2/2	2/2	2/2	1/1
max	106/143	93/121	93/117	256/348

Table 1: Statistics of IWSLT and MERSENNE, including the number of documents (#doc) and sentences (#sent). 'mean', 'min', and 'max' correspond respectively to the average, minimum, or maximum length of sentences, measured in LLAMA tokens.

		EN			FR		
K	rand.	COS	PMI	BM25	COS	PMI	BM25
1	0.52	0.12	0.01	0.22	0.09	0.10	0.32
2	0.52	0.17	0.09	0.26	0.15	0.10	0.30
3	0.52	0.19	0.11	0.28	0.21	0.11	0.31
4	0.52	0.23	0.10	0.28	0.23	0.11	0.34
5	0.51	0.24	0.10	0.29	0.25	0.12	0.34
6	0.51	0.25	0.10	0.29	0.25	0.13	0.36

Table 2: Extraction error rate on MERGEDTED using COS, PMI, and BM25, for K from 1 to 6, computed on the source (EN, left) or target texts (FR, right).

K	rand.	COS	PMI	BM25
1	0.03	0.39	0.54	0.43
2	0.05	0.47	0.52	0.53
3	0.07	0.60	0.46	0.56
4	0.08	0.61	0.46	0.58
5	0.09	0.63	0.55	0.62
6	0.11	0.64	0.61	0.62

Table 3: Cover rate on MERGEDTED using COS, PMI, and BM25, for K from 1 to 6.

not contain IWSLT nor parallel articles from MERSENNE in their pre-training data. Our experimental pipelines relies on vLLM (Kwon et al., 2023), an efficient framework for text generation. Decoding is performed with a beam width of 5 and a maximum number of new tokens of 256. To determine the impact of K (the maximum number of selected contexts), we vary K from 0 to 6. Regarding context selection, we compute PMI using LLAMA. For the Indep-K baseline, we generate 6 examples in the style of TED talks again using LLAMA. More details regarding the experimental setup are in Appendix B.

4.3 Metrics for DLMT

To evaluate general translation quality, we primarily rely on COMET (Rei et al., 2022) and its document-level variant (d-COMET) (Vernikos et al., 2022), with the reference-based model `wmt22-comet-da`. We also report BLEU[9] (Papineni et al., 2002) and SLIDE (Raunak et al., 2024) with `wmt22-cometkiwi-da` with a window size of 8 sentences and a stride of 6. For lexical consistency, we compute Lexical Translation Consistency Ratio (LTCR) (Lyu et al., 2021; Wang et al., 2025b) for proper nouns annotated using spaCy[10] and aligned using awesome-align (Dou and Neubig, 2021). We also conduct case studies to examine the effectiveness of SELF-RAMT.

5 Examining Context Selection Strategies

In this section, we aim to answer the following questions: (a) how effective are the context selection scores to identify relevant contexts? and (b) how similar are the retrieved segments when retrieval is performed in the source text or in the target text? As discussed above, generating coherent texts ideally requires taking the target context into account. As only the source text is initially available, it is important to verify that source-based retrieval is a reliable substitute for target-based retrieval, simulated using the oracle reference.

Method To assess the context selection criteria for their sensitivity to coherence, we challenge their ability to distinguish sentences extracted from the same documents from other noise segments. Starting with a set of document pairs (X^1, X^2) both containing n sentences in the same language, we randomly shuffle sentences from X^1 with those of X^2, resulting in a combined document $X^{1,2}$. We then retrieve, from the first $2n-1$ segments of $X^{1,2}$, the K most relevant segments for the last sentence ($\mathbf{x}_{2n}$), using each relevance score, and compute the extraction error rate r, defined as the proportion of selected sentences that do not belong to the same document as $\mathbf{x}_{2n}$. For a set of N documents $\{X_l^{1,2}, l = 1 \dots N\}$, from which we retrieve the K context sentences $\{c_{l,1}, \dots, c_{l,K}\}$ for each $\mathbf{x}_{2n}$, r is computed as follows:

$$r = \frac{\sum_{l=1}^{N} \sum_{j=1}^{K} \mathbb{I}(\mathrm{doc}(c_j) \neq \mathrm{doc}(\mathbf{x}_{2n}))}{N \times K},$$

where $\mathrm{doc}(\mathrm{c})$ returns the document index of its input segment c.

[9] We use SacreBLEU (Post, 2018) with signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0. We use the default zh tokenizer for translations into Chinese.

[10] https://spacy.io/usage

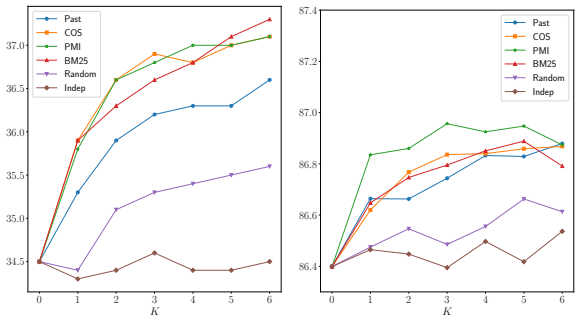


Figure 2: BLEU (left) and s-COMET (right) scores for IWSLT EN–FR translated using EuroLLM, with source and reference as contexts, for K from 0 to 6.

To compare the relevance scores computed using source and target texts, we also compute Kendall's τ (Kendall, 1938, 1945) between the relevance ranking respectively induced on the source and target texts, and average over the N documents. Finally, we also report the cover rate, defined as the ratio of context sentences c_j recognized by retrieval using both $X^{1,2}$ and $Y^{1,2}$, which is the reference translation of $X^{1,2}$, among all selected contexts retrieved from $Y^{1,2}$.[11]

MergedTED We artificially construct shuffled documents $X^{1,2}$ and their translations $Y^{1,2}$ from the 12 EN–FR talks from IWSLT. We consider the first 30 sentences of each talk as a pseudo-document, and include all 66 possible pairs as (X^1, X^2).[12] We refer to the resulting corpus as MergedTED.

Results For question (a), Table 2 reports the extraction error rate for COS, PMI and BM25, derived from MergedTED. We observe that, with the exception of PMI, error rates are quite high: already for $K = 1$, about 12% (resp. 22%) of the sentences retrieved by COS (resp. BM25) do not belong to the same talk as the focus sentence. In comparison, PMI error rates increase more slowly with the value of K. Regarding (b), we note that the error rates computed in the target language (FR) are only slightly higher than in the source (EN). Table 3 reports the cover rates of contextual segments selected in $X^{1,2}$ and in $Y^{1,2}$. Using the source document, the context criteria identify around half of the relevant contexts determined with the reference target document. Regarding the ranking of potential contexts, Kendall's τ between the relevance scores derived from the source and the reference are 0.55, 0.45 and 0.55 for COS, PMI and BM25 respectively. These results provide an empirical support of our main hypotheses, in particular they confirm that we can effectively perform context selection by only looking at the source side of the input documents, yet identify relevant dependencies on the target side.

[11] For the cover rate, we only count segments appearing in the same document as $\mathbf{x}_{2n}$.

[12] We consider 10 different shuffled versions for each pair (X^1, X^2), then report the average r and Kendall's τ.

6 Results and Analyses

6.1 MT quality with SELF-RAMT

The Impact of K To determine the optimal value for K and the best relevance criteria, we examined the BLEU and s-COMET scores for K from 0 (i.e. Zero-shot) to 6, with pairs of source and reference as contexts. Figure 2 displays representative results with the EN–FR translation of IWSLT using EuroLLM, where the context-augmented translations perform better than Random-K and Indep-K. Furthermore, compared to s-COMET, BLEU scores distinguish better translations using selected contexts from Past-K, indicating that these contexts lead to greater lexical similarity between the translations and the references. For a trade-off between quality and complexity, we take $K = 3$ for the following experiments and analysis.

Comparing Relevance Scores We perform context-aware evaluations on IWSLT, reported in Table 4. The results show that all reference-based metrics, including BLEU, s-COMET, and d-COMET, classify PMI as the best or the second-best relevance score for all models and translation directions. In contrast, SLIDE scores are less conclusive, ranking PMI as the top-2 best systems 7 out of 9 times. This suggests that PMI performs better than COS and BM25. LTCR only prefers PMI for EN–FR, while for EN–DE and EN–ZH, baseline methods give higher scores. This highlights a small issue with this metric, when we use the oracle reference as context. Assume that the MT engine translates the first instance of a term x as y_1, different from the reference version (y_2); then, for all the subsequent instances of the same term, we may retrieve the translation of the first instance (y_2), making the system more inclined to generate the same translation (y_2), which is what we want. This will introduce a discrepancy between the first instance (y_1) and the remaining ones, which will be penalized by LTCR. By comparison, the baseline system may appear more consistent.

		EuroLLM					Llama					Qwen				
		BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR
DE	0-shot	27.8	85.4	75.8	81.8	95.8	24.7	82.7	72.3	79.5	95.2	22.8	81.6	70.6	77.2	91.2
	indep	28.1	85.4	75.8	81.8	**96.0**	24.4	82.7	71.9	79.5	95.9	22.8	81.2	69.7	76.5	91.8
	rand.	28.9	*85.7	*76.3	81.8	95.9	25.1	*83.3	*73.0	*79.9	**96.7**	22.6	*82.1	71.1	77.7	**92.8**
	past	29.7	*86.0	*76.8	$^{\Delta}$**82.0**	95.8	26.0	*83.7	*73.4	79.9	93.6	24.0	*82.6	*71.7	77.7	90.3
	COS	**30.2**	*86.1	*76.9	81.8	93.3	**26.8**	*83.9	*73.7	*80.0	94.5	24.4	*83.1	*72.5	**78.0**	88.8
	PMI	30.1	***86.2**	***77.1**	81.8	94.9	26.7	***84.2**	***74.1**	***80.3**	92.6	**25.1**	***83.2**	***72.8**	77.8	87.1
	BM25	29.5	*86.1	*76.8	81.8	93.6	26.6	*83.8	*73.6	79.8	94.9	24.9	*83.0	*72.4	77.9	89.5
FR	0-shot	40.1	86.4	76.8	83.3	88.9	36.5	84.4	74.0	81.8	87.2	34.5	83.9	73.2	80.7	89.6
	indep	41.3	86.4	77.0	83.4	89.9	36.5	84.3	74.0	81.8	87.5	34.8	84.0	*73.6	80.9	88.7
	rand.	41.5	86.5	*77.2	83.3	88.0	37.2	*84.8	*74.5	82.0	88.7	35.2	*84.4	*73.9	*81.2	89.6
	past	42.4	*86.7	*77.6	**83.5**	90.5	38.0	*84.9	*74.8	82.0	85.5	36.5	*84.5	*74.3	*81.4	88.4
	COS	42.8	*86.8	*77.7	83.3	89.7	38.6	*85.0	*75.0	$^{\delta}$**82.1**	86.8	36.7	*84.5	*74.3	$^{\Delta}$80.8	90.0
	PMI	**43.2**	***87.0**	***77.9**	**83.5**	**91.0**	38.6	***85.2**	***75.3**	$^{\delta}$**82.1**	**90.7**	37.1	*84.9	*74.8	*81.4	**93.9**
	BM25	43.1	*86.8	*77.7	83.4	90.4	**39.0**	*85.1	*75.2	82.0	87.9	**37.4**	***85.0**	***74.9**	***81.6**	90.8
ZH	0-shot	30.1	84.4	73.3	**81.3**	75.5	28.3	83.2	70.6	79.2	75.9	29.2	83.2	71.6	78.6	79.0
	indep	30.5	*84.7	*73.8	**81.3**	78.9	29.1	*83.6	*71.9	***79.7**	**76.4**	29.8	*83.7	*72.5	*79.5	81.3
	rand.	30.8	*84.7	*74.0	81.2	78.2	29.4	*83.5	*71.6	79.1	76.3	30.3	*84.1	*73.0	*79.4	**81.5**
	past	31.6	*85.0	*74.4	81.2	**79.7**	30.6	*83.9	*72.4	79.3	74.7	31.5	*84.6	*73.7	*79.4	78.9
	COS	32.0	*84.9	*74.3	80.3	73.2	31.4	*83.9	*72.5	79.2	73.3	32.0	*84.5	*73.7	*79.5	75.9
	PMI	**32.4**	***85.2**	***74.7**	81.0	74.0	**31.6**	***84.0**	***72.9**	79.3	75.1	**32.2**	***84.7**	***74.1**	***79.7**	75.1
	BM25	32.3	*85.0	*74.4	80.6	73.8	31.5	*83.9	*72.4	78.9	72.0	32.1	*84.5	*73.6	*79.5	73.7

Table 4: Results for IWSLT (source and reference as context). We mark MT systems that are significantly better than zero-shot in COMET-based scores, for sentences excluding the first 20 ones ($^{\Delta}$), all sentences ($^{\delta}$), or in both cases (*), with p-value < 0.05. The top two scores are marked in bold (best) and underlined (second-best).

Context		BLEU	s-comet	d-comet	slide	LTCR
	0-shot	55.7	89.5	86.6	74.6	**92.8**
SRC+REF	rand.	58.7	*89.6	*86.8	*74.8	91.7
	past	59.8	*89.8	*87.0	***74.9**	91.8
	COS	60.8	*89.9	*87.2	***74.9**	90.9
	PMI	**61.0**	***90.0**	***87.3**	***74.9**	91.5
	BM25	60.5	*89.9	*87.2	***74.9**	91.2
SRC+MT	rand.	55.8	89.5	86.7	*74.7	91.4
	past	55.0	***89.6**	*86.8	*74.8	92.3
	COS	56.2	***89.6**	***86.9**	*74.8	92.6
	PMI	**56.3**	***89.6**	***86.9**	***74.9**	**92.8**
	BM25	56.2	***89.6**	***86.9**	*74.8	92.7

Table 5: Results for MERSENNE using EuroLLM. * indicates significant gains as in Table 4.

SRC+MT		BLEU	s-comet	d-comet	slide	LTCR
DE	0-shot	27.8	85.4	75.8	81.8	95.8
	rand.	28.5	*85.7	*76.2	82.0	96.0
	past	28.5	*85.7	*76.4	***82.1**	**96.6**
	COS	28.3	*85.7	*76.2	82.0	96.5
	PMI	**28.9**	***85.9**	***76.6**	***82.1**	95.4
	BM25	28.3	*85.7	*76.3	81.9	95.2
FR	0-shot	40.1	86.4	76.8	83.3	88.9
	rand.	41.2	86.5	*77.2	**83.4**	89.2
	past	41.2	86.5	*77.3	**83.4**	87.6
	COS	41.2	86.4	*77.2	**83.4**	92.1
	PMI	**41.8**	***86.6**	***77.5**	**83.4**	90.9
	BM25	41.5	86.4	*77.2	**83.4**	**92.4**
ZH	0-shot	30.1	84.4	73.3	**81.3**	75.5
	rand.	30.6	*84.6	*73.7	81.2	83.1
	past	30.7	84.6	*73.6	81.2	85.1
	COS	30.7	84.5	*73.6	81.2	85.5
	PMI	**30.8**	***84.7**	***73.9**	81.2	**90.0**
	BM25	30.2	84.2	73.2	80.1	88.8

Table 6: Results for IWSLT (source and MT as contexts) using EuroLLM. * indicates significant gains as in Table 4.

Note that this problem does not arise when we retrieve automatic translations, where we see the benefits of SELF-RAMT more clearly.

The evaluation results for MERSENNE and IWSLT translated with reference and automatic translations as context (using EuroLLM) are given in Tables 5 and 6. These results lead to the same conclusion that PMI is a good criterion for retrieving relevant past segments. As the LTCR scores rank the context selection methods differently across test sets, we conduct a follow-up case study presented in Section 6.3, which better highlights the contribution of selected contexts to term consistency. The complete scores for all MT systems are in Tables 11 and 12 in Appendix C.

6.2 Analysis of Context Contribution

Distance Contexts are Retrieved Contextual information plays a crucial role in DLMT. To analyze the contribution of relevant contexts (especially distant ones) to the translation quality, we bin sentences according to the distance (in number of sentences) to their most distant selected context. For example, all selected contexts of sentences in the group 0–20 are retrieved from the past 20 sentences, while there are contexts more distant than 64 sentences for members of the group 64–256.

We then measure the ratio between the effective number of contexts occurring within a given interval (e.g., $20 - 40$) and the maximal possible number of contexts in that interval.[13] This analysis

[13]Which depends on the sentence position in a document: sentences in the initial paragraphs only have access to a re-

Range	rand. nb	rand. ratio	COS nb	COS ratio	PMI nb	PMI ratio	BM25 nb	BM25 ratio
0-20	2066	0.26	3322	0.41	4575	0.57	3432	0.42
20-40	2069	0.26	1708	0.21	1407	0.17	1568	0.19
40-80	2685	0.34	2028	0.25	1417	0.18	2116	0.27
80-120	900	0.21	740	0.17	498	0.12	663	0.16
120-256	356	0.20	278	0.16	179	0.10	297	0.17

Table 7: Distance between the translated sentence and selected contexts, for sentences appearing after the 40^{th} sentence in IWSLT with $K = 3$. 'ratio' denotes the effective value ('nb') normalized by the number of selected contexts for sentences that have access to the corresponding distance interval.

IWSLT Range	DE	FR	ZH
0-20	337	390	373
20-40	241	350	352
40-64	125	223	236
64-256	235	252	258

MERSENNE Range	FR
0-20	966
20-40	770
40-64	650
64-128	995
128-320	570

Table 8: Retrieval statistics with respect to the distance between the translated sentence and selected contexts, for sentences appearing after the 20^{th} sentence. Selection is performed with PMI and $K = 3$, for IWSLT (left) and MERSENNE (right).

is performed for the TED talks test set (IWSLT), for $K = 3$. We exclude from the analysis the first 40 sentences in each document, as the context they can access is limited. The corresponding statistics are in Table 7.

We observe that about half of the retrieved contexts are in the past 20 sentences, while a sizable portion of contexts are chosen in more distant part of the document (in the $20 - 80$ range); more remote sentences, with a distance larger than 80 are also frequently selected. There is a clear variance between relevance scores: COS retrieves closer segments on average, whereas PMI and BM25 are more likely to extract more remote sentences.

Distant Contexts Matter For each group of sentences, we now compare the translations generated with PMI and with the baseline methods. Table 8 reports the corresponding retrieval statistics for these experiments, where we again group sentences by their position index in the source text.

Translation scores are in Table 9, where we compute the d-COMET difference between baselines and PMI-based retrieval. The upper part of Table 9 shows that, for translations of IWSLT with

stricted contexts, while sentences occurring in the last position have a much larger set of contextual segments to chose from.

SRC	The hemihedria is, moreover, non-superposable.
REF	L'hémiédrie est, en outre, non superposable.
0-shot	La hemihedrie est, en outre, non superposable.
Past	De plus, l'hémimorphie n'est pas superposable.
PMI	De plus, l'hémiédrie n'est pas superposable.

Figure 3: Translations of the 72^{nd} sentence of a MERSENNE article, using EUROLLM with MT as target-side context. The correct translations of "hemihedria" are underlined.

SRC: And he found a match.
REF: 结果找到了一个配对！ *(a match)*
0-shot: 他找到了合适的人。*(an appropriate person)*
Past: 他找到了匹配的物品。*(a matched object)*
PMI: 他找到了匹配的物种。*(a matched species)*

The contexts of PMI:
English: It turns out that different species have slightly different structures of collagen, so if you get a collagen profile of an unknown bone, you can compare it to those of known species, and, who knows, maybe you get a match.
Chinese: 事实证明，不同物种的胶原蛋白结构略有不同，因此如果你得到一块未知骨头的胶原蛋白谱，你可以将其与已知物种的胶原蛋白谱进行比较，谁知道，也许能找到匹配。
English: So she shipped him one of the fragments, FedEx.
Chinese: 于是她通过联邦快递把其中一块碎片寄给了他。
English: LN: And he processed it, and compared it to 37 known and modern-day mammal species.
Chinese: LN: 他对它进行了处理，并将其与 37 种已知和现代哺乳动物进行了比较。

Figure 4: EN–ZH translations of the 45^{th} sentence from an IWSLT talk, using EUROLLM with MT as target-side context. PMI retrieves relevant contexts for "a match" that corresponds to a specie (see Figure 7 in Appendix C for more details.)

source and reference as contexts, integrating remote contexts selected by PMI leads to better translation quality than Zero-shot, Random and Past. The bottom part reports the results for translations with source and automatic translations as contexts. In this setting, translation using EUROLLM with PMI is still better than Zero-shot, Random and Past, with lesser performance gains. In contrast, for QWEN, PMI appears to do less well than Past. The performance of LLAMA depends on the language pair, PMI being best only for EN–ZH translations.

We report a similar analysis in Table 10 for MERSENNE corpus. The results show that PMI outperforms Zero-shot and Random for all models using reference or automatic translations as target side context. In all cases except four, PMI is better than Past in d-COMET, especially for QWEN.

6.3 A Case study: Lexical Consistency

We illustrate the benefits of retrieving relevant contexts for the adequate translation in context-dependent cases. A first example is in Figure 4: to translate *a match*, which corresponds here to a

	dist	DE EURO	DE LLAMA	DE QWEN	FR EURO	FR LLAMA	FR QWEN	ZH EURO	ZH LLAMA	ZH QWEN
					SRC+REF					
PMI − sent	0-20	*1.4	*2.3	*2.7	*1.0	*1.2	*1.7	*1.6	*2.8	*2.7
	20-40	*1.4	*1.4	*1.5	*1.3	*1.6	*1.7	*1.5	*2.2	*2.2
	40-64	*1.4	*2.2	*2.1	*0.9	*1.5	*1.9	*1.5	*2.9	*2.9
	64-256	*1.5	*2.2	*2.4	*1.3	*1.9	*1.6	*1.1	*2.0	*3.1
PMI − rand	0-20	*1.1	*1.4	*2.0	*0.7	*1.3	0.5	*0.9	*2.1	*1.6
	20-40	*0.9	0.7	*1.3	*1.0	*0.7	*1.0	*0.9	*1.5	*0.6
	40-64	0.3	*1.2	*2.9	0.5	0.6	*1.0	*1.1	*1.2	*1.5
	64-256	*1.2	*1.7	*1.8	*1.0	*0.8	0.3	0.5	*1.3	*1.5
PMI − past	0-20	0.3	*0.8	0.8	0.2	*0.9	0.3	0.2	0.4	0.1
	20-40	0.2	*0.7	0.7	*0.5	0.5	*0.7	0.4	*0.8	0.3
	40-64	0.3	0.5	*1.8	0.5	*0.7	*1.5	0.4	0.5	*1.0
	64-256	0.6	*1.3	*1.9	0.4	0.4	-0.4	-0.2	*0.9	*0.8
					SRC+MT					
PMI − sent	0-20	*0.7	0.2	0.5	*0.5	-0.2	0.4	*0.7	*1.5	*1.6
	20-40	*0.9	-0.2	-0.7	*0.8	-0.3	0.6	*0.7	*1.2	*0.9
	40-64	*1.6	-0.3	0.8	0.4	0.2	0.7	0.4	*1.3	*1.1
	64-256	*0.8	-0.7	-0.3	*0.8	*0.7	*1.2	*0.7	*1.1	*1.7
PMI − rand	0-20	0.5	0.3	0.4	0.3	0.2	0.3	0.1	*1.2	*0.8
	20-40	*0.7	-0.5	-0.7	*0.5	-0.5	*0.7	*0.6	*1.3	-0.1
	40-64	0.3	-0.6	*2.2	0.3	-0.3	-0.2	0.3	0.5	0.1
	64-256	0.6	-0.5	-0.2	0.3	0.2	0.2	0.4	*1.0	0.6
PMI − past	0-20	0.1	-0.5	-0.8	0.1	-0.1	-0.2	0.2	0.5	-0.1
	20-40	-0.0	-0.4	*-1.0	*0.3	-0.5	0.2	0.4	*0.7	-0.3
	40-64	*1.0	*-0.9	0.6	0.4	0.3	0.7	0.4	0.2	-0.4
	64-256	0.1	*-0.8	-0.2	-0.1	-0.0	-0.1	-0.0	*0.8	0.5

Table 9: Average differences between the d-COMET of translations using PMI and three baselines. From top to bottom these are: zero-shot (sent), random (rand), past (past). We only consider all sentences occurring past the 20^{th} sentence in IWSLT articles, and take into account the maximum distance between the selected contexts and the current sentence. * marks a significant difference with p-value < 0.05.

	dist	SRC+REF EURO	SRC+REF LLAMA	SRC+REF QWEN	SRC+MT EURO	SRC+MT LLAMA	SRC+MT QWEN
sent	0-20	*0.6	*0.9	*1.4	*0.3	*0.5	*0.8
	20-40	*0.7	*0.8	*1.6	*0.4	*0.3	*0.9
	40-64	*0.9	*0.7	*1.9	*0.5	0.3	*1.2
	64-120	*0.7	*0.9	*1.5	*0.2	*0.3	*0.8
	120-320	*0.6	*1.0	*1.6	-0.0	*0.4	*0.6
rand	0-20	*0.3	*0.5	*0.8	0.0	*0.2	*0.3
	20-40	*0.3	*0.5	*0.5	*0.2	*0.3	*0.3
	40-64	*0.6	*0.5	*0.9	*0.3	0.2	*0.5
	64-120	*0.5	*0.6	*0.7	0.1	0.0	*0.3
	120-320	*0.4	*0.9	*0.7	-0.0	0.2	-0.1
past	0-20	-0.0	*0.1	*0.5	-0.0	0.0	*0.3
	20-40	*0.2	*0.2	*0.5	0.1	0.0	*0.2
	40-64	*0.4	0.1	*0.6	0.1	*-0.3	*0.4
	64-120	*0.3	*0.4	*0.5	0.1	0.1	*0.2
	120-320	0.2	*0.4	*1.0	-0.1	0.0	0.4

Table 10: Average differences between the d-COMET of translations using PMI and three baselines. From top to bottom these are: zero-shot (sent), random (rand), past (past). We only consider all sentences occurring past the 20^{th} sentence in MERSENNE articles, and take into account the maximum distance between the selected contexts and the current sentence. * marks significant difference with p-value < 0.05.

matched species, PMI retrieves the relevant contexts and generates the correct translation, while Zero-shot translates it as "an appropriate person" and Past considers it as a matched object. Complementary details and another Chinese example are presented in Appendix C.

Distant contexts are necessary to ensure lexical consistency, especially when translating full articles, and we do observe some evidence for improved translation consistency. For example, in one particular MERSENNE article, the term *hemihedria* appears 12 times,[14] with a consistent French reference translation *hémiédrie*. Table 3 shows the translation of the 72^{nd} sentence using EUROLLM taking source and MT as context.

PMI achieves consistent translation for this term, while Past fails due to the absence of remote contexts, resulting in 11 translation errors with 4 variants. Zero-shot translates them into 6 different forms with only one correct translation, despite its high LTCR score.

Figure 8 and Figure 9 in Appendix C provide the contexts selected using PMI for the 72^{nd} and the 159^{th} sentences respectively, including the sentences in lines 51 and 58 that contain the term *hemihedria*.

7 Conclusion

To achieve better document-level machine translation for extra long documents, we propose the SELF-RAMT framework. By retrieving relevant sentences in a local translation memory composed

[14] Lines 26, 39, 51, 53, 56, 58, 72, 78, 89, 122, 159, 160.

of sentences that have already been translated, and possibly post-edited, in the same document, we expect to generate translations that are globally more consistent. We carry out experiments to translate full TED talks and a novel parallel corpus consisting of scientific articles, using three LLMs with in context learning, to integrate the retrieved past contexts.

We further analyze the influence of distant contexts on translation quality, according to their distance from the current sentence. Our findings show that incorporating distant contexts, selected using context criteria such as PMI, can be useful for better lexical consistency. Distant context also seem to improve the general translation quality, as measured by reference-based scores.

Limitations

In this work, we only considered three open-weight models, excluding close-source models such as GPT4 (OpenAI et al., 2023) the use of which makes results difficult, if possible at all, to reproduce for scientific purposes. The translation of low-resource languages is not included in our experiments, although our framework could be beneficial in that scenario. In our experimental settings, we process each document sentence per sentence, defining the translation unit and the contextually relevant sentences based on this predefined segmentation. Segmenting and processing input documents in larger chunks, made of several consecutive sentences, is left for future work. Our evaluation was based mainly on derivatives of BLEU and COMET metrics, with a variant of LTCR and some case studies. While we reckon that some better metrics should be applied to better reflect document translation quality, measuring for example, the consistency and coherence of the translated text, we also can only regret that such standard evaluation metrics have not yet developed nor adopted by the community.

Ethics Statement

This work has conducted experiments and analysis using open source models, tools, and corpus. We see no ethical problem with this study.

Acknowledgments

This work was supported by the French national agency (ANR) as part of the MaTOS project under reference ANR-22-CE23-0033.[15] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors wish to thank Célia Vaudaine and Caroline Rossi for giving access to the MERSENNE corpus. The authors are also grateful to the anonymous reviewers for their insightful comments and suggestions, and to Paul Lerner and Lichao Zhu for their review and feedback on a preliminary draft of this work.

[15] http://anr-matos.github.io/

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Maxime Bouthors, Josep Crego, and François Yvon. 2024. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3022–3039, Mexico City, Mexico. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine

translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, page 1–31.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Xin Cheng, Shen Gao, Lemao Liu, Dongyan Zhao, and Rui Yan. 2022. Neural machine translation with contrastive translation memories. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3591–3601, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Menglong Cui, Jiangcun Du, Shaolin Zhu, and Deyi Xiong. 2024. Efficiently exploring large language models for document-level machine translation with in-context learning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10885–10897, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yichen Dong, Xinglin Lyu, Junhui Li, Daimeng Wei, Min Zhang, Shimin Tao, and Hao Yang. 2025. Two intermediate translations are better than one: Fine-tuning LLMs for document-level translation refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14917–14933, Vienna, Austria. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Harritxu Gete, Thierry Etchegoyhen, David Ponce, Gorka Labaka, Nora Aranberri, Ander Corral, Xabier Saralegi, Igor Ellakuria, and Maite Martin. 2022. TANDO: A corpus for document-level machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3026–3037, Marseille, France. European Language Resources Association.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and Bobbie Chern et al. 2024. The Llama 3 Herd of Models. Preprint arXiv:2407.21783.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiaxin Guo, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025. Doc-Guided Sent2Sent++: A Sent2Sent++ Agent with Doc-Guided memory for Document-level Machine Translation. Preprint arXiv:2501.08523.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 639–649, Mexico City, Mexico. Association for Computational Linguistics.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. Preprint arXiv:2302.09210.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Martin Kay. 1997. The proper place of men and machines in language translation. *Machine Translation*, 12(3–23).

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Maurice G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bei Li, Hui Liu, Ziyang Wang, Yufan Jiang, Tong Xiao, Jingbo Zhu, Tongran Liu, and Changliang Li. 2020. Does multi-encoder help? a case study on context-aware neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3512–3518, Online. Association for Computational Linguistics.

Chen Li, Meishan Zhang, Xuebo Liu, Zhaocong Li, Derek Wong, and Min Zhang. 2024. Towards demonstration-aware large language models for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13868–13881, Bangkok, Thailand. Association for Computational Linguistics.

Yachao Li, Junhui Li, Jing Jiang, Shimin Tao, Hao Yang, and Min Zhang. 2023. P-transformer: Towards better document-to-document neural machine translation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 31:3859–3870.

Zongyao Li, Zhiqiang Rao, Hengchao Shang, Jiaxin Guo, Shaojun Li, Daimeng Wei, and Hao Yang. 2025. Enhancing large language models for document-level translation post-editing using monolingual data. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8830–8840, Abu Dhabi, UAE. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022a. Divide and rule: Effective pre-training for context-aware multi-encoder translation models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4557–4572, Dublin, Ireland. Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022b. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yuanhua Lv and ChengXiang Zhai. 2011. When documents are very long, bm25 fails! In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, page 1103–1104, New York, NY, USA. Association for Computing Machinery.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring. Preprint arXiv:2407.03618.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Manuel Faysse, Pierre Colombo, François Yvon,

Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical Report. Preprint arXiv:2506.04079.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. Preprint arXiv:2409.16235.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2).

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A lightweight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, and Lenny Bogdonoff et al. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland. European Association for Machine Translation.

Minh Quang Pham, Jitao Xu, Josep Crego, François Yvon, and Jean Senellart. 2020. Priming neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 516–527, Online. Association for Computational Linguistics.

José Pombal, Sweta Agrawal, Patrick Fernandes, Emmanouil Zaranis, and André F. T. Martins. 2024. A context-aware framework for translation-mediated conversations. Preprint arXiv:2412.04205.

Andrei Popescu-Belis. 2019. Context in neural machine translation: A review of models and evaluations. Preprint arXiv:1901.09115.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, and Rui Men et al. 2025. Qwen2.5 technical report. Preprint arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. OpenAI blog.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In *Proceedings of the 2024 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 205–211, Mexico City, Mexico. Association for Computational Linguistics.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Ohad Rubin and Jonathan Berant. 2024. Retrieval-pretrained transformer: Long-range language modeling with self-retrieval. *Transactions of the Association for Computational Linguistics*, 12:1197–1213.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2023. Efficient Transformers: A Survey. *ACM Computing Surveys*, 55(6):1–28.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019b. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Kuang-Da Wang, Teng-Ruei Chen, Yu Heng Hung, Shuoyang Ding, Yueh-Hua Wu, Yu-Chiang Frank Wang, Chao-Han Huck Yang, Wen-Chih Peng, and Ping-Chun Hsieh. 2025a. Plan2align: Predictive planning based test-time preference alignment in paragraph-level machine translation. Preprint arXiv:2502.20795.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025b. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware

data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Mengzhou Xia, Guoping Huang, Lemao Liu, and Shuming Shi. 2019. Graph based translation memory for neural machine translation. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Jitao Xu, Josep Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online. Association for Computational Linguistics.

Jian Yang, Yuwei Yin, Shuming Ma, Liqun Yang, Hongcheng Guo, Haoyang Huang, Dongdong Zhang, Yutao Zeng, Zhoujun Li, and Furu Wei. 2023. Hanoit: Enhancing context-aware translation via selective context. In *Database Systems for Advanced Applications: 28th International Conference, DASFAA 2023, Tianjin, China, April 17–20, 2023, Proceedings, Part III*, page 471–486, Berlin, Heidelberg. Springer-Verlag.

Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, and Jie Zhou. 2019. Enhancing context modeling with a query-guided capsule network for document-level translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1527–1537, Hong Kong, China. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with Bayes' rule. *Transactions of the Association for Computational Linguistics*, 8:346–360.

Emmanouil Zaranis, Nuno M. Guerreiro, and Andre Martins. 2024. Analyzing context contributions in LLM-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, Miami, Florida, USA. Association for Computational Linguistics.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2021. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Zhang Zhuocheng, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556, Singapore. Association for Computational Linguistics.

Álvaro Peris and Francisco Casacuberta. 2019. Online learning for effort reduction in interactive neural machine translation. *Computer Speech & Language*, 58:98–126.

A Mersenne

Due to the scarcity of complete parallel scholarly documents, we constructed MERSENNE, which consists of 23 articles in English and their translation into French prepared by the Mersenne Center.[16] Nineteen of them report recent research in the geosciences domain and the remaining four belong to the chemical sciences. The translations are human post-edits of an initial machine-translated version.

For each article, we first convert the curated html page to plain texts using `pandoc`.[17] Extra empty lines and the symbol xa0 are removed before normalizing the texts to the NFC[18] format through `unicodedata`.[19] We then extract the article from the processed plain text, excluding equations and tables, which are reserved for future exploitation. We segment the texts into sentences and align the sentences to parallel articles. For sentence segmentation, we use Trankit (Nguyen et al., 2021) as it recognizes lists of citations well. Our alignment tool is derived from BertAlign (Liu and Zhu, 2022), which supports many-to-many alignments. All the alignments matched with an empty string were checked and manually adjusted whenever needed.

[16] https://www.centre-mersenne.org/

[17] https://pandoc.org/

[18] Normalization Form Canonical Composition.

[19] https://docs.python.org/3/library/unicodedata.html

We subsequently evaluated the aligned sentence pairs using TransQuest (Ranasinghe et al., 2020). All alignment scores were above 0.75, suggesting that sentence alignment is mostly correct and that all aligned sentences can be kept. Statistics regarding the MERSENNE parallel articles can be found in Table 1.

B Experimental Details

This section presents details about the inference settings, the computation of relevance scores, and the prompt patterns for ICL.

B.1 Inference

In our experiments, we obtain automatic translations through ICL using vLLM (Kwon et al., 2023) in bfloat16. We set the beam width to 5 and the temperature to 0. The minimum and maximum number of new tokens are 1 and 256 respectively.

For the decoding process, we propose a multi-turn decoding algorithm to access the incrementally generated target-side contexts. This involves translating the i^{th} sentences of all documents from the same batch in parallel, and updating the pre-selected contexts for the $(i+1)^{th}$ sentences with the generated translations. When contexts comprise source texts and reference translations of past sentences, we integrate them in the LLM prompts and decode them all at once.

Regarding context selection, we compute PMI scores for source texts using LLAMA, performing compuations in bfloat32 for IWSLT and in bfloat16 for MERSENNE and MERGEDTED.

B.2 Implementation of BM25L

In practice, given a document X, we preprocess each sentence,[20] then compute the term frequency and the inverse document frequency (IDF) for each term in the whole document. We also precomputed an IDF for terms from the training split of IWSLT-2016 (Cettolo et al., 2012)[21] for the EN–FR language pair. Therefore, the IDF of a term in X is replaced by the precomputed values if available.

B.3 Prompt Patterns

We integrate the bilingual context for DLMT in a few-shot template for ICL. As all our MT engines are instruction-tuned, we use the chat template. Regarding prompt design, we empirically tested the performance of the Past-K baseline using different prompt patterns on the IWSLT TST2010 and TST2011 test sets, for $K \in \{2, 6\}$. After disregarding some patterns that lead to over-generation, we selected the prompt templates in Figure 6, where we integrate the few-shot contexts into system prompts for EUROLLM and LLAMA, and into user prompts for QWEN.

For each sentence, we apply the prompt pattern without context to perform zero-shot translation. For the Indep-K baseline, we maintain K-shot demonstrations for all sentences. In practice, we generate 6 examples in the style of TED talks using LLAMA,[22] then integrate the first K examples as K-shot demonstrations.

C Complementary Evaluation Results

Translation Quality Based on the complete evaluation scores for IWSLT with source and automatic translations as contexts (see Table 11), and MERSENNE (see Table 12), we can confirm that in general PMI performs better than COS and BM25, as discussed in Section 6.1. On the other hand, we also observed that the performance of LLAMA is worse than EUROLLM. This quality degradation has a strong influence on DLMT using multi-turn decoding.

Case Study In this section, we provide additional examples that illustrate the need to integrate remote contexts for term consistency when translating long documents, and relevance criteria such as PMI can identify such contexts. This analysis is complementary to the one in Section 6.3. For example, a consistent translation of the term *hemihedria* in the 159^{th} sentence of a MERSENNE article requires contexts more distant than the past 30 sentences, and contexts with the K highest PMI scores include these relevant but remote sentences (see Figure 9). Figure 5 displays another example that contextual information from the past 10^{th} sentence is required for the consistent translation of "the High Arctic".

Figure 7 gives complementary information for the example in Figure 4, including the two successive sentences of the source sentence to be trans-

[20] The preprocessing consists of lowercasing, stop-word removal and stemming using PyStemmer: https://pypi.org/project/PyStemmer/.

[21] https://wit3.fbk.eu/2016-01

[22] We use the following chat template integrated in the system prompt "You are a good translation assistant!" and the user prompt "Give six few-shot examples to assist the translation from English to {tgt_lang} in the style of TED talks."

		EuroLLM					Llama					Qwen				
		BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR
DE	0-shot	27.8	85.40	75.77	81.83	95.85	24.7	82.73	72.25	79.52	95.23	22.8	81.64	70.61	77.21	91.16
	indep	28.1	85.42	75.81	81.83	96.02	24.4	82.69	71.91	79.46	95.88	22.8	81.23	69.71	76.54	91.83
	rand.	28.5	*85.70	*76.18	81.98	96.01	24.4	82.88	72.41	79.73	95.08	22.7	81.75	70.65	77.43	93.48
	past	28.5	*85.69	*76.43	***82.07**	**96.57**	24.5	*83.09	*72.71	***80.08**	96.64	**23.0**	81.93	71.07	77.74	91.61
	COS	28.3	*85.69	*76.24	82.00	96.50	24.5	82.85	72.27	79.71	**97.95**	22.9	***82.15**	$^{\delta}$**71.19**	77.81	**95.72**
	PMI	**28.9**	***85.92**	***76.62**	*82.06	95.38	24.2	82.68	72.16	*79.97	96.04	22.3	81.69	70.75	77.72	91.58
	BM25	28.3	*85.72	*76.32	81.94	95.24	**24.9**	***83.16**	***72.76**	79.72	97.82	22.9	81.72	70.74	**77.90**	93.75
FR	0-shot	40.1	86.40	76.84	83.34	88.89	36.5	84.36	73.98	81.78	87.20	34.5	83.87	73.20	80.68	89.64
	indep	41.3	86.39	77.03	83.37	89.89	36.5	84.28	73.97	81.77	87.53	34.8	84.03	*73.58	80.88	88.66
	rand.	41.2	86.48	*77.18	83.41	89.21	36.5	**84.40**	74.03	81.79	88.81	34.7	84.02	73.45	*81.17	89.91
	past	41.2	86.52	*77.32	83.43	87.65	**36.7**	84.32	74.08	**81.93**	88.23	35.0	83.97	*73.71	$^{\delta}$81.12	89.46
	COS	41.2	86.43	*77.16	83.43	92.14	**36.7**	84.23	73.89	**81.93**	92.53	35.1	84.07	73.52	$^{\Delta}$81.05	93.35
	PMI	**41.8**	***86.64**	***77.53**	**83.45**	90.87	**36.7**	84.27	73.99	81.84	**93.63**	**35.2**	84.07	*73.85	81.03	**94.71**
	BM25	41.5	86.43	*77.18	83.43	**92.43**	36.6	84.29	**74.12**	81.73	92.22	35.1	**84.17**	***73.91**	***81.35**	94.38
ZH	0-shot	30.1	84.42	73.31	81.28	75.52	28.3	83.21	70.62	79.17	75.88	29.2	83.24	71.56	78.63	79.02
	indep	30.5	*84.65	*73.76	**81.30**	78.91	**29.1**	***83.59**	***71.89**	***79.70**	76.40	29.8	*83.73	*72.53	*79.54	81.34
	rand.	30.6	*84.62	*73.65	81.22	83.08	26.6	83.10	70.87	79.14	78.70	29.9	*83.90	*72.49	*79.39	79.44
	past	30.7	84.60	*73.64	81.18	85.12	26.9	83.20	*71.33	79.30	82.89	30.1	***84.02**	***72.98**	***79.64**	87.26
	COS	30.7	84.49	*73.56	81.20	85.50	27.1	82.98	*71.04	78.96	81.92	30.0	*83.93	*72.69	*79.55	**89.94**
	PMI	**30.8**	***84.69**	***73.86**	81.23	**90.02**	27.3	$^{\Delta}$83.46	*71.86	79.35	**86.14**	**30.3**	*84.00	*72.90	*79.57	86.11
	BM25	30.2	84.21	73.23	80.15	88.79	26.8	83.12	*71.23	79.14	85.02	**30.3**	*83.87	*72.59	*79.22	88.37

Table 11: Evaluation for the translation of IWSLT, translated using source and automatic translation as context, for $K = 3$. We mark significantly positive difference between context-augmented methods and zero-shot MT for COMET-based scores, for sentences excluding the first 20 ones ($^{\Delta}$), all sentences ($^{\delta}$), or in both cases (*), with p-value < 0.05.

		EuroLLM					Llama					Qwen				
		BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR	BLEU	s-comet	d-comet	slide	LTCR
REF	0-shot	55.7	89.48	86.64	74.60	**92.85**	51.0	88.41	85.34	73.20	**92.59**	47.6	87.58	84.18	71.99	90.12
	rand.	58.7	*89.60	*86.78	*74.82	91.73	52.6	*88.65	*85.62	*73.39	92.09	51.8	*88.25	*85.01	*72.93	89.79
	past	59.8	*89.80	*87.02	*74.88	91.79	53.8	*88.89	*85.89	***73.52**	91.28	53.1	*88.37	*85.10	*72.91	**90.74**
	COS	60.8	*89.94	*87.22	*74.88	90.89	55.0	*89.03	*86.04	***73.52**	91.85	55.4	*88.68	*85.54	***73.33**	90.06
	PMI	**61.0**	***89.97**	***87.26**	***74.91**	91.54	**55.2**	***89.14**	***86.16**	*73.51	92.32	**55.8**	*88.80	*85.67	*73.26	90.68
	BM25	60.5	*89.91	*87.17	*74.90	91.21	54.9	*89.11	*86.09	*73.44	91.32	55.4	***88.82**	***85.70**	*73.24	89.98
MT	0-shot	55.7	89.48	86.64	74.60	**92.85**	51.0	88.41	85.34	73.20	92.59	47.6	87.58	84.18	71.99	90.12
	rand.	55.8	89.51	86.72	*74.75	91.45	51.5	*88.55	*85.52	73.31	92.87	48.4	*88.01	*84.74	*72.81	90.63
	past	55.0	*89.56	*86.84	*74.85	92.34	51.5	*88.63	***85.69**	***73.57**	92.89	48.3	*87.91	*84.74	*72.97	92.09
	COS	56.2	*89.59	*86.87	*74.82	92.64	51.5	*88.61	*85.61	*73.44	93.62	48.7	*88.04	*84.90	*73.06	92.62
	PMI	**56.3**	*89.61	***86.90**	***74.88**	92.77	**51.6**	*88.64	*85.67	*73.56	**93.96**	**49.0**	***88.09**	***85.01**	***73.16**	92.79
	BM25	56.2	***89.63**	***86.90**	*74.85	92.69	**51.6**	***88.66**	*85.66	*73.45	93.94	48.5	*88.02	*84.89	*73.06	**92.92**

Table 12: Evaluation for the translation of MERSENNE, translated using bilingual contexts with source and reference (REF, top) or source and automatic translation (MT, bottom), for $K = 3$. We mark significantly positive difference between context-augmented methods and zero-shot MT for COMET-based scores, for sentences excluding the first 20 ones ($^{\Delta}$), all sentences ($^{\delta}$), or in both cases (*), with p-value < 0.05.

<|im_start|>system
Translate the following English source text to Chinese, considering the provided English context and its Chinese translations:
English: that Natalia had dug out of the High Arctic belonged to ...
Chinese: Natalia 从加拿大北极高地挖出的东西属于……
English: So this camel would have been about nine feet tall, weighed around a ton.
Chinese: 因此，这只骆驼身高约 9 英尺，重约 1 吨。
English: But chances are the postcard image you have in your brain is one of these, the dromedary, quintessential desert creature -- hangs out in sandy, hot places like the Middle East and the Sahara, has a big old hump on its back for storing water for those long desert treks, has big, broad feet to help it tromp over sand dunes.
Chinese: 但是，很可能你脑海中浮现的明信片图像是其中一种，即单峰驼，这种动物是典型的沙漠生物，生活在中东和撒哈拉等炎热干燥的地方，背上有一大块驼峰，用来储存水分以应对漫长的沙漠探险，还有大而宽的脚，帮助它在沙丘上行走。<|im_end|>
<|im_start|>user
English: So how on earth would one of these guys end up in the High Arctic?
Chinese:<|im_end|>
<|im_start|>assistant

Figure 5: Selected contexts using PMI for the 56^{th} sentence from a IWSLT talk for EN–ZH translation. It included the 46^{th} sentence that contains the first mention of "the High Arctic".

lated ("And he found a match"), and also the contexts used in Past and PMI, with $K = 3$. The contexts selected using PMI, stating that the person can "compare it to those of known species" and "get a match", produce an adequate translation of "a match".

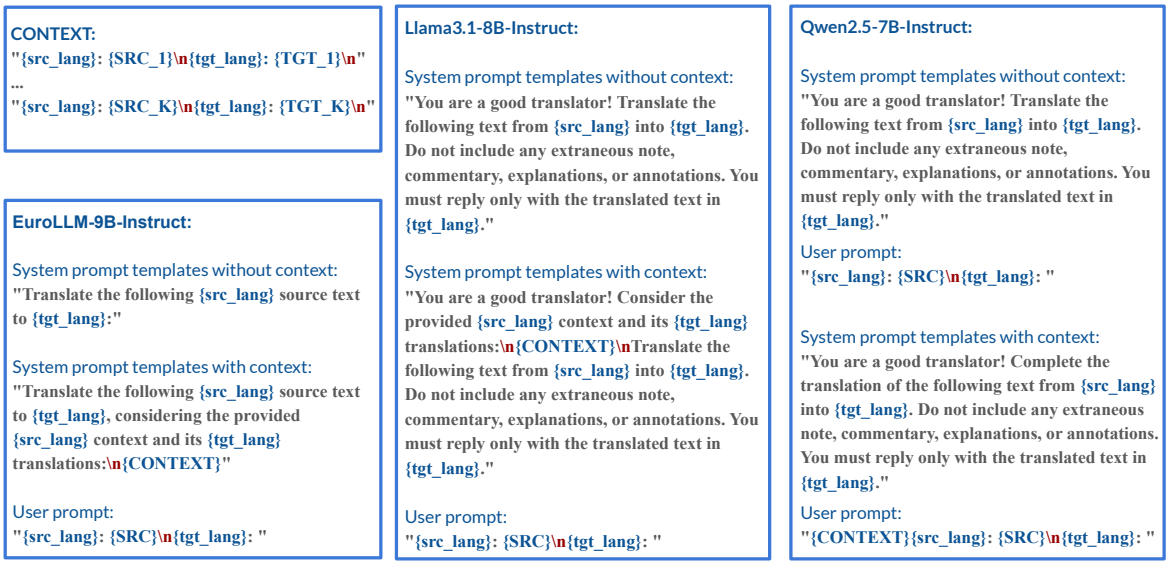


Figure 6: The prompt patterns for EUROLLM, LLAMA and QWEN applied in our experiments.

SRC: #1 And he found a match. #2 that Natalia had dug out of the High Arctic belonged to ... #3 a camel. **0-shot:** #1 他找到了合适的人。*(an appropriate person)* #2 纳塔莉亚从北极地区挖出来的东西属于…… #3 骆驼 **Past:** #1 他找到了匹配的物品。*(a matched object)* #2 纳塔莉亚从高北极地区挖出来的东西属于…… #3 一只骆驼。 **PMI**: #1 他找到了匹配的物种。*(a matched species)* #2 Natalia 从加拿大北极高地挖出的东西属于…… #3 一只骆驼。	**The contexts of Past for #1:** **English**: So she shipped him one of the fragments, FedEx. **Chinese**: 于是她用联邦快递把碎片寄给了他。 **English**: NR: Yeah, you want to track it. It's kind of important. **Chinese**: NR：是的，你想追踪它。这很重要。 **English**: LN: And he processed it, and compared it to 37 known and modern-day mammal species. **Chinese**: LN：然后他进行了处理，并将其与37种已知和现代哺乳动物进行了比较。	**The contexts of PMI for #1:** **English**: It turns out that different species have slightly different structures of collagen, so if you get a collagen profile of an unknown bone, you can compare it to those of known species, and, who knows, maybe you get a match. **Chinese**: 事实证明，不同物种的胶原蛋白结构略有不同，因此如果你得到一块未知骨头的胶原蛋白谱，你可以将其与已知物种的胶原蛋白谱进行比较，谁知道，也许能找到匹配。 **English**: So she shipped him one of the fragments, FedEx. **Chinese**: 于是她通过联邦快递把其中一块碎片寄给了他。 **English**: LN: And he processed it, and compared it to 37 known and modern-day mammal species. **Chinese**: LN：他对它进行了处理，并将其与 37 种已知和现代哺乳动物进行了比较。

Figure 7: EN–ZH translations of the 45^{th} sentence from an IWSLT talk, using EUROLLM with MT as target-side context. PMI retrieves relevant contexts in the 41^{th} sentence for "a match", which means a matched species.

<|im_start|>system
Translate the following English source text to French, considering the provided English context and its French translations:
English: In the memoir of May 22, 1848[4], it is the "tartrates" and the "paratartrates" which are primarily considered, but the young scientist seeks fruitful generalisations: "It will be said, and rightly so: All organic substances that deviate from the plane of polarisation when they are dissolved will therefore enjoy hemihedria.
French: Dans le mémoire du 22 mai 1848 [4], ce sont les « tartrates » et les « paratartrates » qui sont principalement considérés, mais le jeune scientifique cherche des généralisations fructueuses : « On dira, et à juste titre : toutes les substances organiques qui dévient du plan de polarisation lorsqu'elles sont dissoutes profiteront donc de l'hémiédrie.
English: It was even by studying this latter property that I was assured of the hemihedria, which I then realised via careful observation of the crystalline form.
French: C'est même en étudiant cette dernière propriété que j'ai été convaincu de l'hémiédrie, que j'ai ensuite confirmée par une observation attentive de la forme cristalline.
English: At this stage, Pasteur had moved away from the morphological study of the crystals to the study of the possible rotations of the plane of polarisation which they induced, which had led him to better characterise the hemihedria of tartrates.
French: À ce stade, Pasteur s'était éloigné de l'étude morphologique des cristaux pour étudier les rotations possibles du plan de polarisation qu'ils induisaient, ce qui l'avait conduit à mieux caractériser l'hémiédrie des tartrates.<|im_end|>
<|im_start|>user
English: The hemihedria is, moreover, non-superposable.
French:<|im_end|>
<|im_start|>assistant

Figure 8: Selected contexts for the 72^{nd} sentence of an MERSENNE article using PMI, including the 51^{st}, 56^{th} and 58^{th} sentences containing the term *hemihedria*.

<|im_start|>system
Translate the following English source text to French, considering the provided English context and its French translations:
English: His appointment to the University of Lille, in an industrial environment that led him to study amyl alcohols, helped to reorient his scientific activity, but he remained mainly driven by his hypothesis that “molecular dissymmetry” was the prerogative of the living.
French: Sa nomination à l'université de Lille, dans un environnement industriel qui l'a conduit à étudier les alcools amyles, a contribué à réorienter son activité scientifique, mais il est resté principalement guidé par son hypothèse selon laquelle la « dissymétrie moléculaire » était l'apanage du vivant.
English: In the memoir of May 22, 1848[4], it is the "tartrates" and the "paratartrates" which are primarily considered, but the young scientist seeks fruitful generalisations: "It will be said, and rightly so: All organic substances that deviate from the plane of polarisation when they are dissolved will therefore enjoy hemihedria.
French: Dans le mémoire du 22 mai 1848 [4], ce sont les « tartrates » et les « paratartrates » qui sont principalement considérés, mais le jeune scientifique cherche des généralisations fructueuses : « On dira, et à juste titre : toutes les substances organiques qui dévient du plan de polarisation lorsqu'elles sont dissoutes profiteront donc de l'hémiédrie.
English: At this stage, Pasteur had moved away from the morphological study of the crystals to the study of the possible rotations of the plane of polarisation which they induced, which had led him to better characterise the hemihedria of tartrates.
French: À ce stade, Pasteur s'était éloigné de l'étude morphologique des cristaux pour étudier les rotations possibles du plan de polarisation qu'ils induisaient, ce qui l'avait conduit à mieux caractériser l'hémiédrie des tartrates.<|im_end|>
<|im_start|>user
English: This preceded the second thesis, devoted to amyl alcohols, where Pasteur examined the crystallographic question thus posed, and where he writes to have not succeeded in inducing crystalline hemihedria.
French:<|im_end|>
<|im_start|>assistant

Figure 9: Selected contexts for the 159^{th} sentence of an MERSENNE article using PMI, including the 51^{st} and 58^{th} sentences containing the term *hemihedria*.

Using Encipherment to Isolate Conditions for the Successful Fine-tuning of Massively Multilingual Translation Models

Carter Louchheim, Denis Sotnichenko, Yukina Yamaguchi and Mark Hopkins
Williams College
Williamstown, MA

Abstract

When fine-tuning massively multilingual translation models for low-resource languages, practitioners often include auxiliary languages to improve performance, but factors determining successful auxiliary language selection remain unclear. This paper investigates whether syntactic similarity or lexical overlap is more important for effective multilingual fine-tuning. We use encipherment to create controlled experimental conditions that disentangle these confounded factors, generating novel languages with identical syntax but no lexical overlap, and conversely, languages that preserve lexical overlap. Through extensive NLLB-200 fine-tuning experiments across Europarl and AmericasNLP datasets, we demonstrate that lexical overlap is the dominant factor. Syntactically identical auxiliary languages provide negligible benefits (< 1.0 ChrF), while languages with significant lexical overlap provide substantial improvements (> 5.0 ChrF), with effectiveness strongly correlated to KL-divergence between token distributions ($r = -0.47$, $p < .001$). Our findings provide clear guidance: when selecting auxiliary languages for multilingual fine-tuning, prioritize lexical overlap over syntactic similarity.

1 Introduction

A popular modern approach to low-resource machine translation is the fine-tuning of massively multilingual encoder-decoder transformers (Vaswani et al., 2017). For example, the top two entrants in the AmericasNLP 2023[1] Shared Task on Machine Translation into Indigenous Languages (which solicits systems that translate Spanish into Indigenous American languages) both used this technique (Ebrahimi et al., 2023). The teams from the University of Sheffield (Gow-Smith and Sánchez Villegas, 2023) and the University of Helsinki (De Gibert et al., 2023) fine-tuned distillations of Meta's NLLB-200 model (Costa-Jussà et al., 2022) simultaneously on all eleven language pairs of the shared task.

Simultaneous training on a set of related language pairs (as opposed to training a separate system per language pair) has frequently been reported to yield performance benefits (Aharoni et al., 2019; Maillard et al., 2023). One survey (Ranathunga et al., 2023) on low-resource machine translation claims: "This is mainly due to the capability of the model to learn an interlingua (shared semantic representation between languages)". But when the parent model (as in the case of NLLB-200) has already been pre-trained on 200 language pairs, when (and why) does it remain beneficial to simultaneously fine-tune on multiple language pairs? What do the fine-tuning languages learn from each other?

In the context of multilingual language modeling, investigators have focused on two candidates: syntax and lexicon. According to a recent review (Philippy et al., 2023): "In previous research, syntax has been suggested as potentially the most important linguistic contributor for better cross-lingual transfer."[2] The same article also reports that "lexical overlap is particularly important when the pre-training corpus for the source language is small or when the word order between the source and target languages is dissimilar," but concludes that "lexical overlap is not a sufficient standalone factor to explain cross-lingual transfer."

An obstacle to drawing definitive conclusions is the difficulty of isolating the confounding factors of shared syntax and lexicon – related languages typically share both. In this work, we use **encipherment** to disentangle these factors. Encipher-

[1] These two systems were the baselines for the 2024 edition of the shared task, and continued to be the top systems for most language pairs. The Sheffield system was used as the baseline for the 2025 shared task, and maintained its superiority.

[2] The review, however, hypothesizes that the impact of syntax "may be overestimated" due to shortcomings in the research methods.

Proceedings of the Tenth Conference on Machine Translation, pages 241–252
November 8-9, 2025

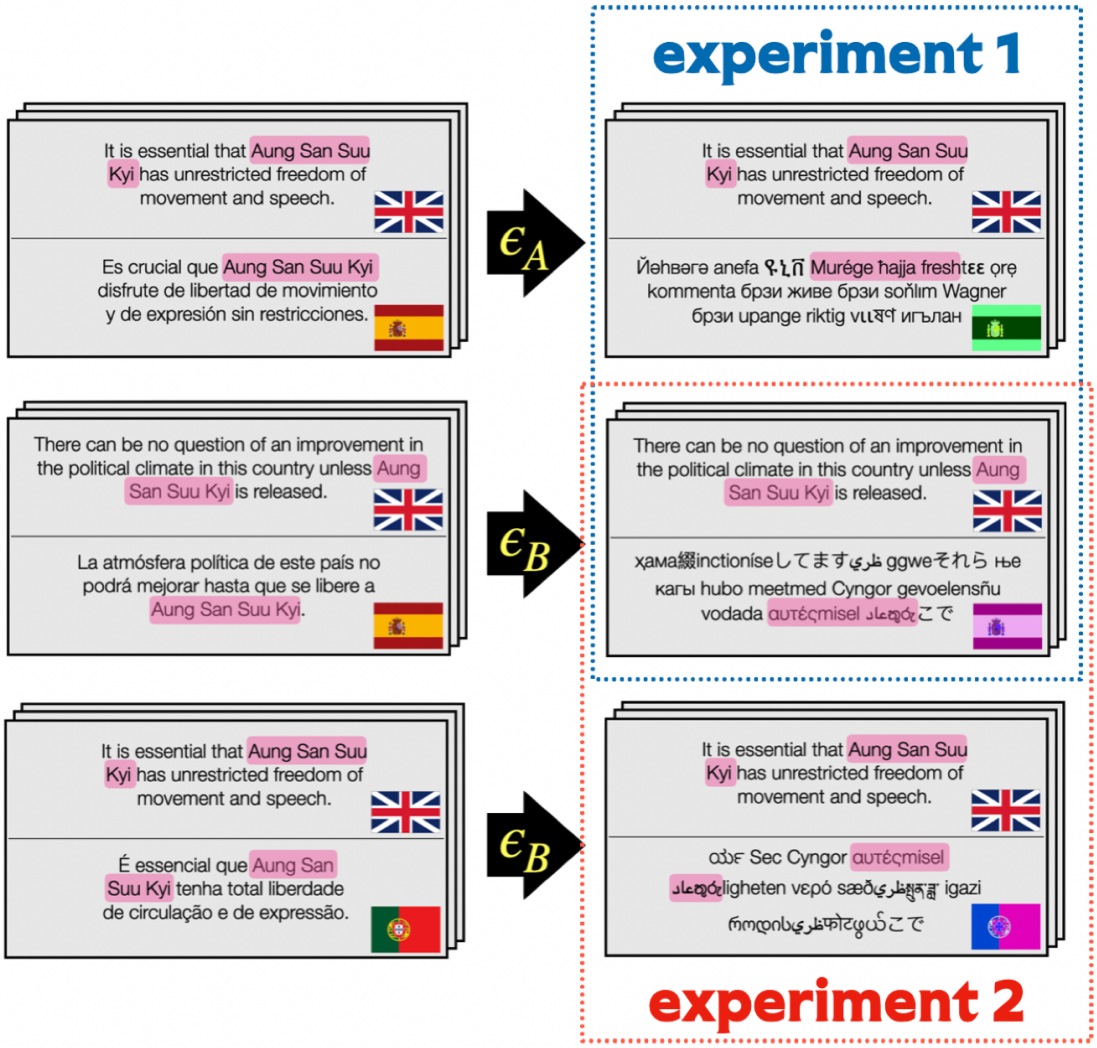


Figure 1: We use encipherment to create statistically realistic languages that are unseen by pre-trained translation models, affording experimental control over the often-confounded factors of syntactic and lexical overlap. For the top two rows, we apply different encipherments to disjoint subsets of English-Spanish Europarl, creating two novel languages with identical syntax but no lexical overlap. For the bottom two rows, we apply the same encipherment to disjoint subsets of English-Spanish and English-Portuguese Europarl, creating two novel languages that preserve the lexical overlap between Spanish and Portuguese.

ment allows us to generate statistically realistic languages that have not been previously included in translation model pre-training, while also providing experimental control:

1. We can produce syntactically identical languages with no lexical overlap. Figure 1 (top two rows) shows an example where two English-Spanish corpora are enciphered using different encipherments (ϵ_A and ϵ_B), producing two languages that are syntactically identical but lexically distinct.

2. We can produce novel languages that preserve cross-lingual lexical overlap. Figure 1 (bottom two rows) also shows an example where an English-Spanish corpus and an English-Portuguese corpus are enciphered using the same encipherment (ϵ_B), producing two languages that preserve the lexical overlap between Spanish and Portuguese.

The goal of this paper is to provide practical guidance to those seeking to build translation engines for low-resource languages. Specifically, when fine-tuning a massively multilingual translation model like NLLB-200, how should one select auxiliary languages to include in the fine-tuning (or should one include them at all)? We ultimately arrive at the following recommendations:

- **Recommendation 1:** Lexical overlap is the

most important factor to consider. If there is a low relative entropy (KL-divergence) between the token distribution of two source or target languages, then you can get considerable performance benefits from multilingual fine-tuning.

- **Recommendation 2:** Even if you can find an auxiliary language with extreme grammatical similarity to your low-resource language of interest, the performance benefits of multilingual fine-tuning (attributable to common syntax) are liable to be negligible.

Bottom line: When choosing auxiliary languages for the multilingual fine-tuning of massively multilingual translation models, focus on languages with high lexical overlap with your low-resource language of interest.

2 Related Work

Neural Machine Translation

Zoph et al. (2016) approached low-resource neural machine translation by leveraging a "parent" model (pre-trained on a high-resource language pair) to train a "child" model (for a low-resource language pair). Nguyen and Chiang (2017) and Kocmi and Bojar (2018) streamlined this process so that it could be succinctly described as follows (Kocmi and Bojar, 2018): "We train the parent language pair for a number of iterations and switch the training corpus to the child language pair for the rest of the training, without resetting any of the training (hyper)parameters."

Ensuing work studied conditions resulting in successful transfer from a parent to a child model. Among this work, Dabre et al. (2017) explored several parent-child combinations and reported that "transfer learning done on a X-Y language pair to [a] Z-Y language pair has maximum impact when Z-Y is resource-scarce and when X and Z fall in the same or linguistically similar language family." Lin et al. (2019) trained gradient-boosted decision tree models to predict synergistic parent/child language pairs, and observed that dataset size and word overlap were the most common splitting features. Aji et al. (2020) determined that the "inner" transformer layers were more crucial to transfer than the embedding layer, and noted that even using a simple copy model as the parent had performance benefits over training from scratch. This earlier work focused on parent models that were trained on a single language pair.

Over the past few years, the trend has been to pre-train massively multilingual translation models (Aharoni et al., 2019; Costa-Jussà et al., 2022) by simultaneously training on many language pairs. Focusing on the pragmatics of this training paradigm, Shaham et al. (2023) studied "interference," i.e. when multilingual pre-training underperforms bilingual pre-training. They concluded that the main cause of interference is when the model size is too small relative to the available training data. Tan and Monz (2023) used a linear regression model to determine factors that predict the zero-shot X→Y translation performance (where neither X nor Y is English) of multilingual models trained exclusively on English-centered language pairs. Our focus is on fine-tuning massively multilingual parent models for a low-resource language pair, specifically the factors that promote synergistic multilingual fine-tuning. We also believe we are the first work in this space to use encipherment as an instrument to de-confound the factors of syntactic similarity and lexical overlap.

Multilingual Language Modeling

It was quickly observed (Lin et al., 2019; Pires et al., 2019) that pre-trained multilingual encoder-only language models like mBERT (Devlin et al., 2019) could be fine-tuned on certain tasks (like named entity recognition or textual entailment) using monolingual supervision (typically English supervision), and only suffer minor performance degradation when applied zero-shot to other languages from the pre-training corpus. Several papers have studied this phenomenon (Dufter and Schütze, 2020; K et al., 2020; Lauscher et al., 2020; Ahuja et al., 2022; Deshpande et al., 2022; de Vries et al., 2022; Wu et al., 2023) – enough to have merited a survey paper (Philippy et al., 2023). Among these papers, K et al. (2020) used the most similar methodology to ours. To determine the impact of lexical overlap on cross-lingual transfer, they pre-trained two versions of BERT: one on English and Hindi[3] and another on "fake English" and Hindi, where "fake English" was derived from English by shifting the Unicode encoding of each character by a fixed constant. They concluded that lexical overlap plays only a minor role in cross-lingual transfer for textual entailment and named entity

[3]They also conducted this experiment with Russian instead of Hindi.

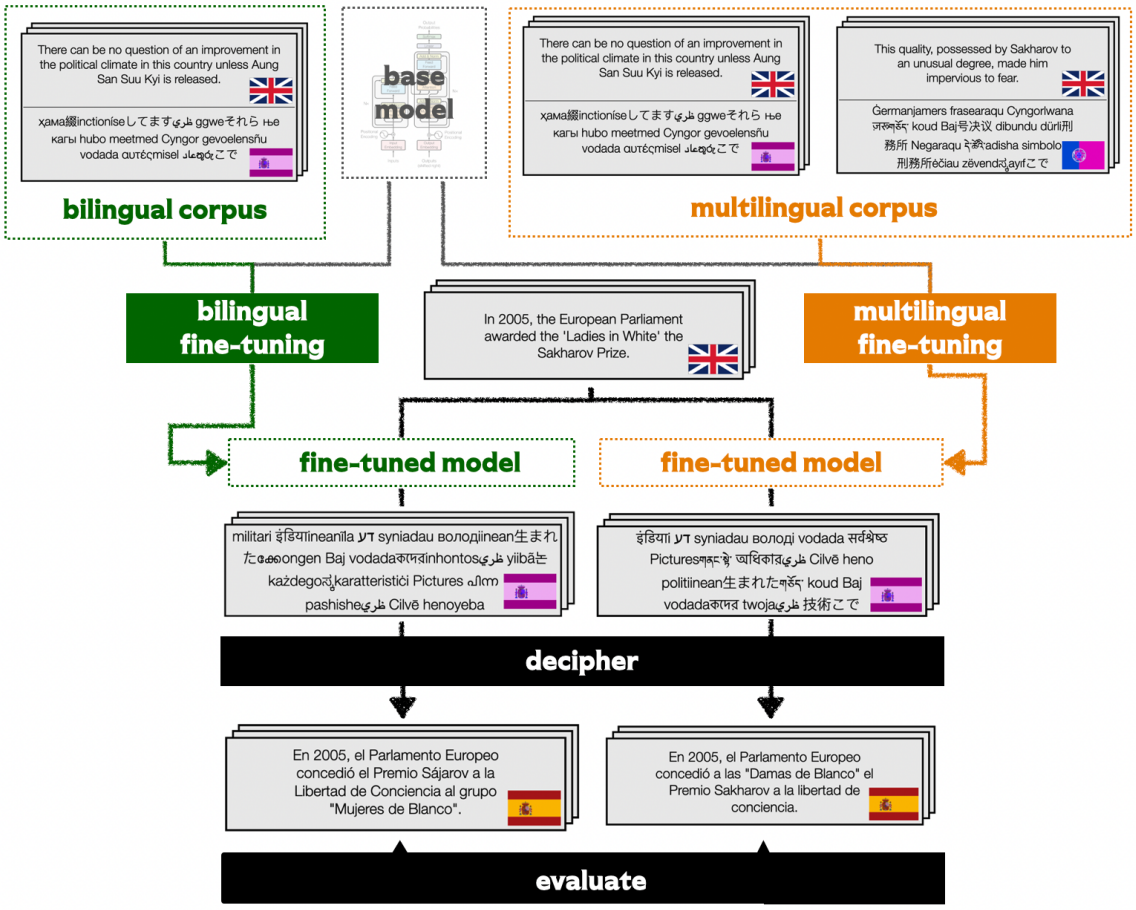


Figure 2: Overview of an experiment trial: a base model is fine-tuned using one or more enciphered bitexts. The resulting model is evaluated by translating held-out test sets, then deciphering and scoring the translations.

recognition. Also relevant to our approach is the work of Wu et al. (2023), who used "controlled studies" to investigate the factors contributing to the success of cross-lingual transfer learning – for instance, they perform artificial syntactic manipulations before fine-tuning on the GLUE dataset (Wang et al., 2018).

3 Preliminaries

This paper uses the following formalisms to describe enciphered parallel corpora.

Let $\mathcal{T}$ and $\mathcal{L}$ be finite alphabets that respectively correspond to a token vocabulary[4] and a set of language ids (e.g. `eng_Latn`, `rus_Cyrl`, etc.). Let $\mathcal{T}^*$ be the set of all sequences of tokens from $\mathcal{T}$.

Define a *parallel corpus* as a function $\pi : D \mapsto \mathcal{T}^*$, where $D \subset \mathcal{L} \times \mathbb{Z}^+$ and $\forall (l_1, i), (l_2, i) \in D$, $\pi(l_1, i)$ and $\pi(l_2, i)$ have the same meaning (i.e. they are translations of one another).

Define an *encipherment* ϵ as a permutation[5] of token vocabulary $\mathcal{T}$, i.e. a bijection $\epsilon : \mathcal{T} \mapsto \mathcal{T}$. For a token sequence $\pi(l, i) = \langle t_1, ..., t_k \rangle$ from parallel corpus π, denote the ϵ-enciphered sequence as $\pi_\epsilon(l, i) = \langle \epsilon(t_1), ..., \epsilon(t_k) \rangle$.

We extract a *bitext* from parallel corpus π using the following notation:

$$\hat{\pi}(l, l', \epsilon, \epsilon', I) = \{(\pi_\epsilon(l, i), \pi_{\epsilon'}(l', i) \mid i \in I\}$$

where $l, l' \in \mathcal{L}$ are language ids, ϵ, ϵ' are encipherments, and $I \subset \mathbb{Z}^+$ is a finite set of indices.

4 Experiment Design

Figure 2 provides an overview of our experiment design. Each experiment involves K bitexts ex-

[4]Throughout this paper, we assume a fixed token vocabulary. Namely, the token vocabulary we use in all our experiments is the NLLB-200 (Costa-Jussà et al., 2022) token vocabulary.

[5]One might worry that a translation model could simply learn to invert the permutation, but previous experimental work (Aji et al., 2020) suggests "that the [transformer] model is incapable of untangling [an] embedding permutation."

tracted from parallel corpus π:

$$\hat{\pi}(l_1, l'_1, \epsilon_1, \epsilon'_1, I_1)$$
$$\hat{\pi}(l_2, l'_2, \epsilon_2, \epsilon'_2, I_2)$$
$$\vdots$$
$$\hat{\pi}(l_K, l'_K, \epsilon_K, \epsilon'_K, I_K)$$

The index sets $I_1, \ldots, I_K$ are pairwise disjoint.

We use these bitexts to train several translation models:

- **bilingual fine-tuning**: For each bitext $\hat{\pi}(l_k, l'_k, \epsilon_k, \epsilon'_k, I_k)$, we fine-tune model M_k from pre-trained model M_{base}.
- **multilingual fine-tuning**: We use the entire collection of bitexts to simultaneously fine-tune a single model M_{multi} from pre-trained model M_{base}. During training, we sample evenly from the bitexts.

We fine-tune each model using a batch size of 64 for a maximum of 60,000 training steps. Validation is performed every 500 steps, and training is terminated early if the validation loss does not decrease for five consecutive evaluations. We use the Adafactor optimizer (Shazeer and Stern, 2018), as implemented in the Transformers library (Wolf et al., 2020), with a fixed learning rate of 1×10^{-4}, disabling both parameter scaling and relative step sizing. Gradient clipping is applied with a threshold of 1.0, and a weight decay of 1×10^{-3} is used for regularization. We adopt a constant learning rate schedule with warm-up, increasing the learning rate linearly over the first 1,000 steps.

We evaluate the resulting models by translating held-out test sets, then deciphering the translations and scoring them using standard machine translation metrics (e.g. BLEU (Papineni et al., 2002) and ChrF (Popović, 2015)). Because one of the languages in our experiments is polysynthetic, we report results using ChrF, but the choice of metric does not affect the experimental conclusions. We run 5 trials for each of the following training bitext sizes: 1024, 2048, 4096, 8192, and 16392 (so 25 trials in total). A bilingual fine-tuning trial consists of a fine-tuning for each bitext – the resulting systems are then evaluated and their scores are averaged. A multilingual fine-tuning trial is a single fine-tuning over all bitexts – the resulting system is then evaluated on each language pair, and these scores are averaged.

5 Base Models

We focus on fine-tuning NLLB-200 models (Costa-Jussà et al., 2022). These models were trained on a large-scale multilingual corpus covering 200 languages, using a transformer-based encoder-decoder architecture, following the general design of the M2M-100 model (Fan et al., 2021). Several distillations of this model are provided, including a 600M parameter model (6 encoder/decoder layers, 768 hidden size, 12 attention heads) and a 1.3B parameter model (12 encoder/decoder layers, 1024 hidden size, 16 attention heads). To increase experimental throughput, our experiments focus on the 600M parameter model.

6 Datasets

This section describes the parallel corpora that we use in our experiments.

Europarl

The Europarl Parallel Corpus (Koehn, 2005) is a set of sentence-aligned proceedings from the European Parliament covering sessions from 1996 to 2011. It spans 21 European languages, with each language contributing approximately 60 million words across 30 million aligned sentence fragments. We preprocess the corpus to eliminate repeated sentences.

AmericasNLP

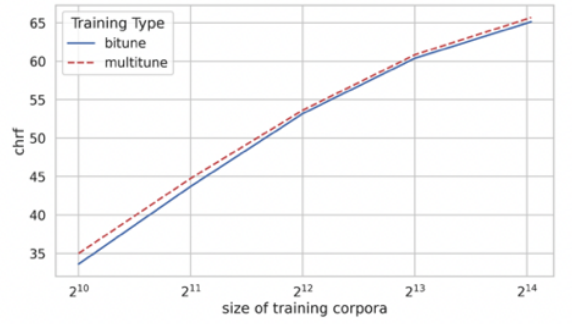
Referenced in the introduction, the AmericasNLP Workshop solicits systems that translate Spanish into Indigenous American languages. They provide official training corpora for these language pairs. In 2025, the workshop introduced Spanish → Wayuunaiki as a new language pair (Prieto et al., 2024). Wayuunaiki is "an Arawakan language spoken in northern Colombia and Venezuela, primarily by the Wayuu community, with about 420,000 speakers. It is an agglutinative language with a predominant SOV word order." (De Gibert et al., 2025)

7 Experiment 1: The Impact of Syntactic Similarity on the Success of Multilingual Fine-tuning

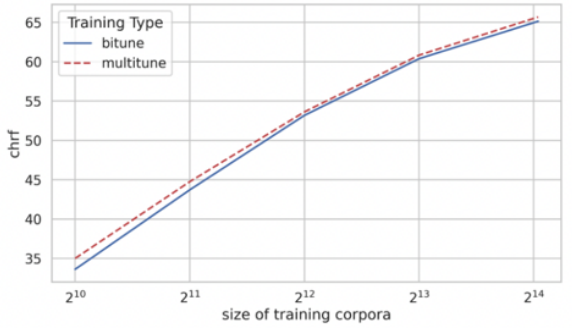
7.1 Syntactic Similarity of Target Languages

For our first set of experiments, we construct a scenario in which we have target languages with no lexical overlap but identical syntax (i.e. the top two rows of Figure 1). Specifically, we extract the

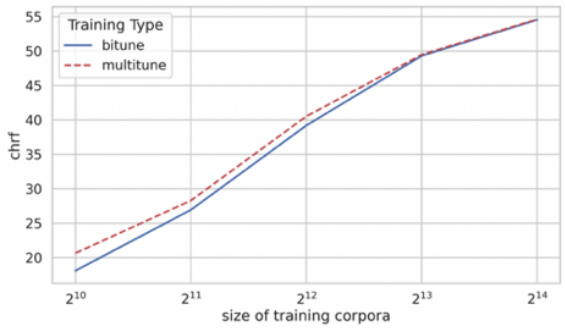
base model	corpus	source	target (enciphered)
nllb-200(600M)	europarl	english	spanish-ε1 spanish-ε2



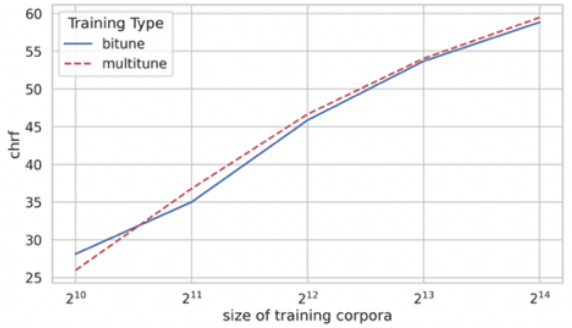
base model	corpus	source (enciphered)	target
nllb-200(600M)	europarl	spanish-ε1 spanish-ε2	english



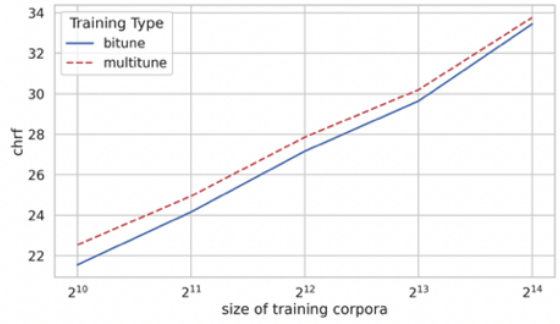
base model	corpus	source	target (enciphered)
nllb-200(600M)	europarl	english	czech-ε1 czech-ε2



base model	corpus	source (enciphered)	target
nllb-200(600M)	europarl	czech-ε1 czech-ε2	english



base model	corpus	source	target (enciphered)
nllb-200(600M)	anlp	spanish	wayuunaiki-ε1 wayuunaiki-ε2



base model	corpus	source (enciphered)	target
nllb-200(600M)	anlp	wayuunaiki-ε1 wayuunaiki-ε2	spanish

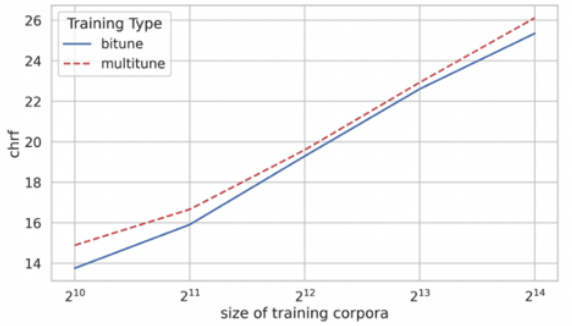


Figure 3: Experiment 1 results. Multilingual fine-tuning using two lexically distinct but syntactically identical languages provides only marginal improvement over bilingual fine-tuning of each language independently. While more pronounced for Wayuunaiki (whose unenciphered analogue is not part of the NLLB-200 pre-training corpus), the benefits are still minor (< 1.0 ChrF).

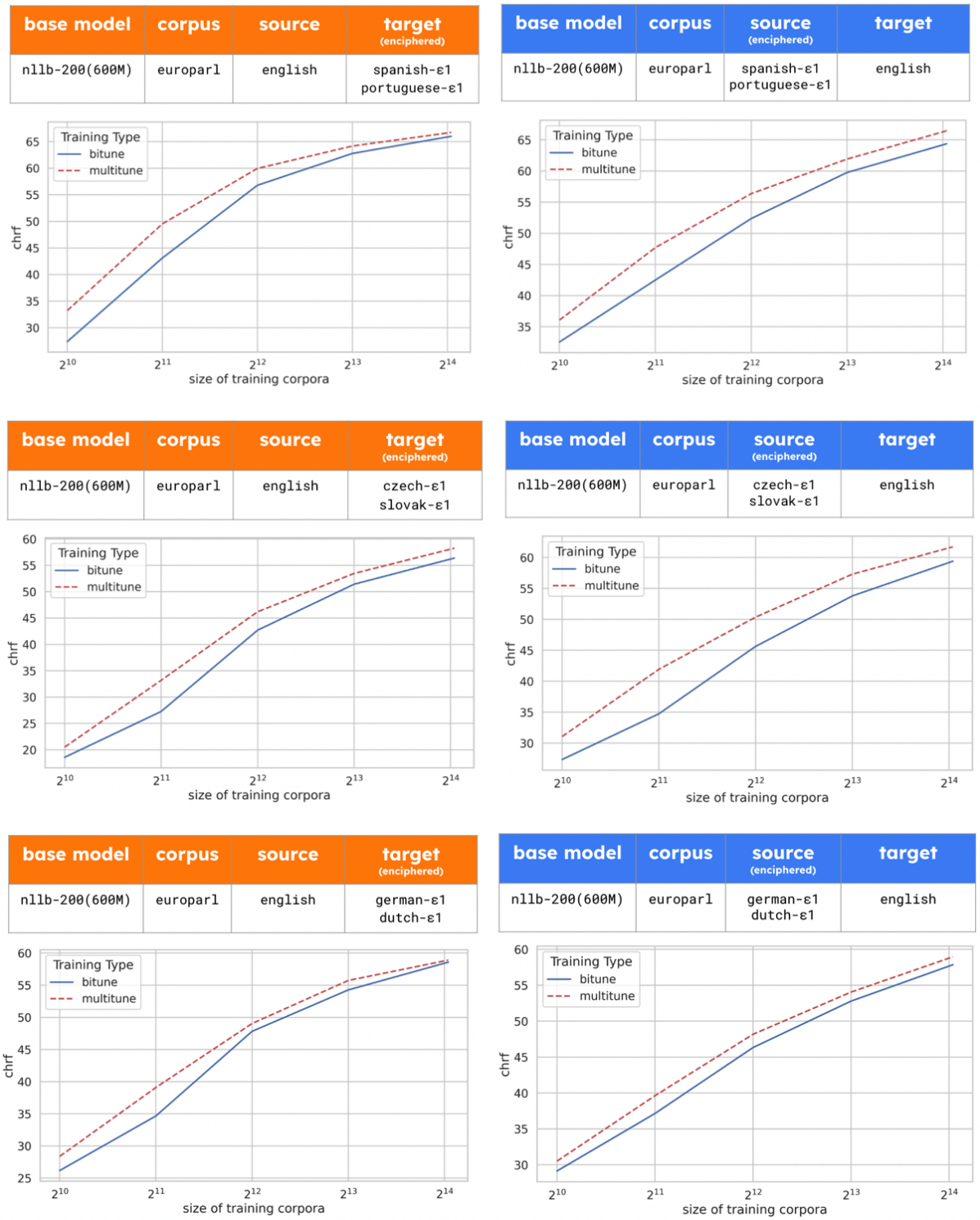


Figure 4: Experiment 2 results. Multilingual fine-tuning using two enciphered languages that preserve lexical overlap can provide significant improvement ($>$ 5.0 ChrF at certain data sizes) over bilingual fine-tuning of each language independently.

following k bitexts from parallel corpus π:

$$\hat{\pi}(l, l', \epsilon^0, \epsilon'_1, I_1)$$
$$\hat{\pi}(l, l', \epsilon^0, \epsilon'_2, I_2)$$
$$\vdots$$
$$\hat{\pi}(l, l', \epsilon^0, \epsilon'_k, I_k)$$

Here (and henceforth), ϵ^0 denotes the identity function (i.e. no encipherment occurs, since the encipherment maps each token to itself). Note that we have constructed k target languages with identical syntax but distinct lexicons. During fine-tuning, we freeze the encoder to eliminate the confounding impact of encoder domain adaptation to the source language.

7.2 Syntactic Similarity of Source Languages

Analogously, we construct a scenario in which we have **source** languages with no lexical overlap but identical syntax, i.e., we extract the following k bitexts from parallel corpus π:

$$\hat{\pi}(l, l', \epsilon_1, \epsilon^0, I_1)$$
$$\hat{\pi}(l, l', \epsilon_2, \epsilon^0, I_2)$$
$$\vdots$$
$$\hat{\pi}(l, l', \epsilon_k, \epsilon^0, I_k)$$

This time, we have constructed k source languages with identical syntax but distinct lexicons. During fine-tuning, we freeze the **decoder** to eliminate the confounding impact of decoder domain adaptation to the target language.

7.3 Results

We conducted these experiments using the following language pairs:

- Target Side Encryption: English $\rightarrow$ {Spanish, Czech, Wayuunaiki}
- Source Side Encryption: {Spanish, Czech, Wayuunaiki} $\rightarrow$ English

Figure 3 shows results from this set of experiments. Multilingual fine-tuning generally shows little advantage over bilingual fine-tuning, even though the two target languages are syntactically identical (they are both encipherments of the same language). Only in the case of Wayuunaiki, a language whose family (Arawak) was not represented in the original NLLB-200 training set, do we observe a small benefit from multilingual fine-tuning. This suggests that even in the best-case scenario – where we can find an auxiliary language with nearly identical syntax to our low-resource language of interest – the benefits of multilingual training (in the absence of lexical overlap) is minor.

Conclusion: Syntactic similarity of the source or target languages appears to have little impact on the effectiveness of multilingual fine-tuning.

8 Experiment 2: The Impact of Lexical Overlap on the Success of Multilingual Fine-tuning

8.1 Lexical Overlap of Target Languages

The only difference between this experiment and Experiment 1 is that we extract the following k bitexts from parallel corpus π:

$$\hat{\pi}(l, l'_1, \epsilon^0, \epsilon', I_1)$$
$$\hat{\pi}(l, l'_2, \epsilon^0, \epsilon', I_2)$$
$$\vdots$$
$$\hat{\pi}(l, l'_k, \epsilon^0, \epsilon', I_k)$$

We use different languages but the same encipherment, so that shared tokens remain shared after encipherment (see the bottom two rows of Figure 1). This produces unseen languages that have the same amount of lexical overlap as their unenciphered analogues. Again, we freeze the encoder during fine-tuning to eliminate the confounding impact of encoder domain adaptation to the source language.

8.2 Lexical Overlap of Source Languages

We also construct a scenario in which we have unseen **source** languages that have the same amount of lexical overlap as their unenciphered analogues, i.e., we extract the following k bitexts from parallel corpus π:

$$\hat{\pi}(l'_1, l, \epsilon', \epsilon^0, I_1)$$
$$\hat{\pi}(l'_2, l, \epsilon', \epsilon^0, I_2)$$
$$\vdots$$
$$\hat{\pi}(l'_k, l, \epsilon', \epsilon^0, I_k)$$

During fine-tuning, we freeze the **decoder** to eliminate the confounding impact of decoder domain adaptation to the target language.

8.3 Results

For these experiments, we used English as the unenciphered language l. As the enciphered

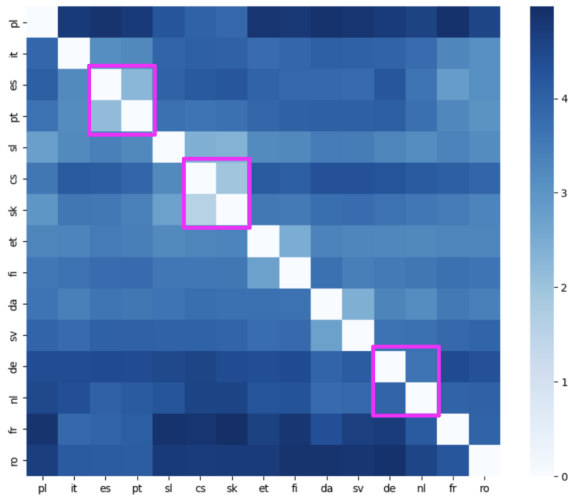


Figure 5: KL-divergence heatmap between Europarl languages. The heatmap shows the KL-divergence between token distributions for all pairs of Latin script languages in the Europarl corpus. Lighter colors indicate lower KL-divergence (greater lexical overlap). The magenta boxes highlight language pairs used in Experiment 2: Spanish-Portuguese (KL $\approx$ 2.26), Czech-Slovak (KL $\approx$ 1.96), and German-Dutch (KL $\approx$ 3.63). This visualization helps explain why Spanish-Portuguese and Czech-Slovak multilingual fine-tuning show greater benefits than German-Dutch, as their lower KL-divergence values indicate higher lexical overlap.

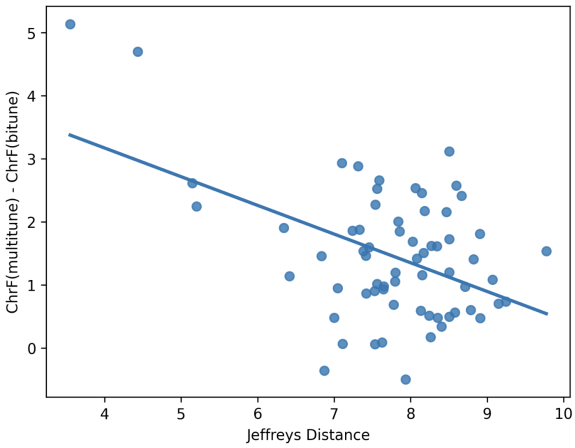


Figure 6: Relationship between lexical overlap and multilingual fine-tuning effectiveness. The scatter plot shows the ChrF improvement (multilingual minus bilingual fine-tuning) versus the Jeffreys distance between token distributions for all pairs of non-English Latin script languages in Europarl. Each point represents a single trial of Experiment 2 with bitext size 4096. The negative correlation (r = -0.47, p < .001) demonstrates that languages with greater lexical overlap (lower Jeffreys distance) benefit more from multilingual fine-tuning, supporting the conclusion that lexical overlap is the primary factor driving successful auxiliary language selection.

language combinations, we used three pairs of geographically-proximate European languages (which we assumed would have significant lexical overlap):

- l'_1 = Spanish and l'_2 = Portuguese
- l'_1 = Czech and l'_2 = Slovak
- l'_1 = German and l'_2 = Dutch

Figure 4 shows the results of these experiments. Under these conditions, multilingual fine-tuning substantially outperforms bilingual fine-tuning, often by more than five ChrF points. Given that Experiment 1 showed little benefit to incorporating syntactically similar languages during multilingual fine-tuning, it would appear that the observed benefits are mainly attributable to the lexical overlap.

However, the multilingual fine-tuning of English $\leftrightarrow$ German and English $\leftrightarrow$ Dutch is notably less effective than the others. To explain this difference, we computed the KL-divergence between the token distributions of all Europarl languages that use Latin script (see Figure 5). The KL-divergence from Spanish to Portuguese (approximately 2.26) and from Czech to Slovak (approximately 1.96) is considerably smaller than the KL-divergence from German to Dutch (approximately 3.63).

To assess the general impact of lexical overlap on the effectiveness of multilingual training, we conducted a single trial of Experiment 2 (using bitext size 4096) for every pair (l'_1, l'_2) of non-English Latin script languages in Europarl. Figure 6 plots the ChrF delta between multilingual and bilingual fine-tuning, versus the Jeffreys distance (additive symmetrization of KL-divergence) between languages l'_1 and l'_2. There is a moderate, statistically significant negative correlation ($r(63) = -0.47$, $p < .001$), suggesting that lexical overlap is a significant determining factor in the effectiveness of multilingual fine-tuning.

9 Conclusion

This work addresses a fundamental question in low-resource machine translation: when fine-tuning massively multilingual models like NLLB-200, which factors determine the success of multilingual fine-tuning with auxiliary languages? Through controlled experiments using encipherment to disen-

tangle syntactic similarity and lexical overlap, we provide empirical evidence that lexical overlap is the primary driver of performance improvements. Our key findings are:

- Syntactic similarity provides minimal benefit: Even when auxiliary languages share identical syntax with the target language, multilingual fine-tuning shows little advantage over bilingual approaches. The benefits are most pronounced (but still minor) only for languages from families not represented in the pre-training corpus.
- Lexical overlap drives substantial improvements: Languages that share vocabulary can provide significant performance gains (> 5.0 ChrF in many cases), with effectiveness inversely correlated to the KL-divergence between token distributions.

The encipherment methodology introduced here also provides an experimental framework for future research on cross-lingual transfer, allowing researchers to control for confounding factors that typically make it difficult to isolate the impact of different linguistic properties.

Future work should explore whether these findings generalize to other massively multilingual models, investigate optimal methods for measuring and maximizing lexical overlap, and examine whether the relative importance of syntax versus lexicon changes with different model architectures or pre-training objectives.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural preservation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages

7–14, Mexico City, Mexico. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*.

Shaomu Tan and Christof Monz. 2023. Towards a better understanding of variations in zero-shot neural machine translation performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13553–13568, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. 2023. Oolong: Investigating what makes transfer learning hard with controlled studies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3280–3289, Singapore. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Translate, then Detect: Leveraging Machine Translation for Cross-Lingual Toxicity Classification

Samuel J. Bell[*†] **Eduardo Sánchez**[*†‡] **David Dale**[†]
Pontus Stenetorp[‡] **Mikel Artetxe**[§] **Marta R. Costa-jussà**[†]
[†]Meta [‡]University College London
[§]University of the Basque Country (UPV/EHU)
{eduardosanchez, alastruey, chrisropers, costajussa}@meta.com
p.stenetorp@cs.ucl.ac.uk mikel.artetxe@ehu.eus

Abstract

Multilingual toxicity detection remains a significant challenge due to the scarcity of training data and resources for many languages. While prior work has leveraged the *translate-test* paradigm to support cross-lingual transfer across a range of classification tasks, the utility of translation in supporting toxicity detection at scale remains unclear. In this work, we conduct a comprehensive comparison of translation-based and language-specific/multilingual classification pipelines. We find that translation-based pipelines consistently outperform out-of-distribution classifiers in 81.3% of cases (13 of 16 languages), with translation benefits strongly correlated with both the resource level of the target language and the quality of the machine translation (MT) system. Our analysis reveals that traditional classifiers outperform large language model (LLM) judges, with this advantage being particularly pronounced for low-resource languages, where `translate-classify` methods dominate `translate-judge` approaches in 6 out of 7 cases. We additionally show that MT-specific fine-tuning on LLMs yields lower refusal rates compared to standard instruction-tuned models, but it can negatively impact toxicity detection accuracy for low-resource languages. These findings offer actionable guidance for practitioners developing scalable multilingual content moderation systems.

1 Introduction

Detecting instances of toxic, abusive, or hateful content at scale is a challenging problem with important, real-world implications for content moderation. In a multilingual setting, however, toxicity detection is often rendered particularly difficult due to a paucity of labeled data for lower-resourced languages. In parallel, recent years have seen the scaling up of machine translation (MT) systems to cover a vast array of world languages (e.g., NLLB-Team et al., 2022), offering a potential pathway toward leveraging cross-lingual transfer for improved multilingual toxicity detection.

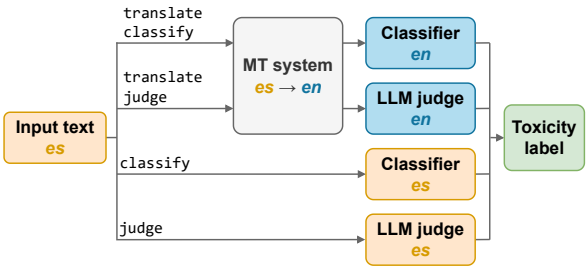


Figure 1: Across 17 languages, we evaluate toxicity detection using translation-based pipelines (`translate-classify`, `translate-judge`) against classifying in the original language (`classify`, `judge`). In this example, text in Spanish (es) is optionally translated to English (en) before classification.

In monolingual non-English settings, cross-lingual transfer has already proven useful for toxicity detection (Eskelinen et al., 2023; Kobellarz and Silva, 2022), aligning with broader analyses of translation's utility for cross-lingual transfer across a range of classification tasks (Artetxe et al., 2023; Etxaniz et al., 2023b; Ponti et al., 2021). Specifically, Artetxe et al. (2023) compare *translate-test* (translating a sample before zero-shot classification) against *translate-train* (translating a sample before classification with a classifier *finetuned on translation data*) and find that *translate-test* is competitive as long as translation quality is sufficient.

While cross-lingual classification has been widely studied in other NLP tasks, toxicity detection presents distinctive challenges that warrant separate investigation. Toxic language is culturally and contextually grounded, with expressions, slurs, and taboos that often lack direct equivalents across languages, making transfer more brittle than for semantically simpler labels. Online toxicity also frequently involves code-switching, orthographic

[*]Joint first author

Proceedings of the Tenth Conference on Machine Translation, pages 253–268
November 8-9, 2025

variation, and deliberate obfuscation, which may be less common in other tasks. Moreover, toxicity labels are inherently subjective and shaped by cultural norms, leading to potential label drift when transferring across languages. These factors, combined with the high stakes of moderation errors, make cross-lingual transfer in toxicity detection both consequential and scientifically challenging.

In this work, we present an empirical exploration of translation for multilingual toxicity detection, through the lens of the practitioner for whom labeled data may be unavailable—a particularly common scenario when working with lower-resourced languages—by comparing translation-based pipelines against a variety of off-the-shelf multi- and monolingual classifiers. Across 27 pipelines spanning five MT systems and nine toxicity classifiers—including both traditional classifiers and large language model (LLM) judges—we evaluate the benefit of cross-lingual classification in 17 languages with varying levels of resources.

Our results suggest that leveraging translation is an effective method for multilingual toxicity detection (§4.1), with benefits scaling in line with increasing language resources and MT system quality (§4.2). Motivated by these results, we study the issue of refusal rates and its mitigation via MT supervised finetuning (MT-SFT), as well as the downstream effect of MT-SFT on toxicity detection performance (§4.3). Finally, we explore classifying using LLM judges and compare them to traditional toxicity classifiers (§4.4). We conclude with practical recommendations for deploying multilingual toxicity detection systems at scale.

2 Related work

2.1 Multilingual toxicity detection

Multilingual toxicity detection is widely used in cases like content moderation or faithful translation (e.g. Costa-jussà et al., 2023). Prior work has either trained models using multilingual corpora of labeled training data (e.g. Hanu, 2020), or sought to exploit cross-lingual transfer via monolingual finetuning of multilingual foundation models (e.g. XLM-ROBERTa; Conneau et al., 2020a). Multilingual evaluation datasets exist for toxicity detection (e.g. Kivlichan et al., 2020; Gupta, 2021) alongside those used for text detoxification (Dementieva et al., 2024b, 2025). In this work, we evaluate a representative sample of off-the-shelf traditional classifiers, including cross-lingual, and both mono- and multilingual classifiers, across a wide variety of languages.

2.2 Cross-lingual classification

Early approaches to cross-lingual classification relied on bilingual lexicons and statistical methods to project documents into a shared feature space (Rapp, 1995; Dumais et al., 1997; Gliozzo and Strapparava, 2006). The introduction of cross-lingual word embeddings (Mikolov et al., 2013; Faruqui and Dyer, 2014; Ammar et al., 2016) enabled models trained in one language to be applied to others through a shared vector space. Prior to multilingual encoders, transfer was typically achieved via MT, either by translating the training data into the target language (*translate-train*) or by translating inputs into the source language at inference (*translate-test*) (Wan, 2009; Prettenhofer and Stein, 2010).

Multilingual sentence encoders such as LASER (Artetxe and Schwenk, 2019) and mBERT (Devlin et al., 2019) demonstrated the feasibility of direct zero-shot transfer without translation. XLM (Lample and Conneau, 2019) introduced translation language modeling to improve alignment, and XLM-R (Conneau et al., 2020b) showed consistent gains from scaling model and data size. Artetxe et al. (2020) provided a systematic comparison of translate-train and translate-test, while Etxaniz et al. (2023a) revisited translate-test with modern neural MT, finding it competitive for low-resource and distant languages.

Recent work explores large multilingual LLMs (Muennighoff et al., 2022) and parameter-efficient adaptation methods (Pfeiffer et al., 2020), aiming to combine the flexibility of fine-tuning with the scalability of zero-shot prompting.

3 Methods

We evaluate the performance of toxicity detection *pipelines*, where a pipeline comprises a binary toxicity classifier and an optional MT system. In many languages—and particularly for lower-resourced languages—labeled data for toxicity detection is unavailable, precluding the training and deployment of specialized classifiers and motivating the consideration of translation-based pipelines. As such, we are principally interested in comparing pipelines in the following three regimes:

`classify (ID)` An untranslated, in-distribution (ID) sample is classified in the source language

Language Code	Language	FineWeb-2 Docs	Dataset	No. Samples
am	Amharic	280,355	Amharic Hate Speech (Ayele et al., 2023)	1,501
ar	Arabic	57,752,149	L-HSAB (Mulki et al., 2019)	5,846
de	German	427,700,394	GermEval 2018 (Wiegand et al., 2018)	3,398
es	Spanish	405,634,303	Jigsaw Multilingual (Kivlichan et al., 2020)	8,438
fr	French	332,646,715	Jigsaw Multilingual (Kivlichan et al., 2020)	10,920
he	Hebrew	13,639,095	OffensiveHebrew (Hamad et al., 2023)	500
hi	Hindi	20,587,135	MACD (Gupta et al., 2022)	6,728
it	Italian	219,117,921	Jigsaw Multilingual (Kivlichan et al., 2020)	8,494
kn	Kannada	2,309,261	MACD (Gupta et al., 2022)	6,587
ml	Malayalam	3,406,035	MACD (Gupta et al., 2022)	5,170
pt	Portuguese	189,851,449	ToLD-Br (Leite et al., 2020)	21,000
ru	Russian	605,468,615	Russian Language Toxic Comments (Belchikov, 2019)	14,412
ta	Tamil	5,450,192	MACD (Gupta et al., 2022)	6,000
te	Telugu	2,811,760	MACD (Gupta et al., 2022)	6,000
th	Thai	35,949,449	Thai Toxicity Tweet Corpus (Sirihattasak et al., 2018)	2,794
tr	Turkish	88,769,907	Jigsaw Multilingual (Kivlichan et al., 2020)	14,000
uk	Ukrainian	47,552,562	TextDetox 2024 (Dementieva et al., 2024a)	5,000

Table 1: Toxicity datasets used per language, including number of samples, and number of documents in FineWeb-2 as a measure of language resourcedness.

using a classifier trained on data from the same distribution (e.g., evaluating a classifier on French social media posts that has been trained on French social media posts).

`classify (OOD)` An untranslated, out-of-distribution (OOD) sample is classified in the source language using a classifier trained on data from a different distribution (e.g., evaluating a classifier on French video comments that has been trained on French social media posts).

`translate-classify` The sample is translated into English using an MT model before being classified in English, using a toxicity classifier that supports English. No evaluated classifiers have been trained on translated data.

While we expect finetuned classifiers to exhibit the strongest performance while operating ID, it is relative to the far more common OOD scenario (i.e., where no suitably finetuned classifier is available to process the source language) that we expect `translate` pipelines to offer significant utility.

3.1 Evaluation

We evaluate various pipeline implementations across several languages and datasets, each of which comprise text samples x_i and gold toxicity labels y_i. Each pipeline, given a sample, produces a continuous score corresponding to toxicity.

Pipeline performance To avoid the need for thresholding, we evaluate pipeline performance via the Area Under the Receiver Operating Characteristic curve (AUC), which provides a continuous measure of how well the pipeline can separate toxic from non-toxic samples. The AUC is defined as:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t)$$

where $\text{TPR}(t)$ and $\text{FPR}(t)$ are the true positive and false positive rates at threshold t.

When comparing pipelines, we typically evaluate the benefit of using one pipeline over another by way of change in AUC. For two pipelines, P_A and P_B,

$$\Delta\text{AUC}(P_A, P_B) = \text{AUC}(P_A) - \text{AUC}(P_B)$$

We evaluate all possible combinations of pipeline and dataset where the supported pipeline language matches the dataset's language.

Language resources We evaluate the role of language resourcefulness on pipeline performance, where we roughly approximate the number of available resources using the amount of documents available in FineWeb2 (Penedo et al., 2025), a large-scale dataset of web text sourced from various CommonCrawl snapshots.

Translation system quality Following standard practice (e.g., Kocmi et al. 2024), we additionally evaluate the quality of translations into English using the CometKiwi-DA-XL (Rei et al., 2023) quality estimation model, evaluated on the BOUQuET (Omnilingual MT Team et al., 2025) dataset.

Classifier	Supported Languages	Base Model	Training Dataset
xlm-r-finetuned-toxic-political-tweets-es	es	XLM-RoBERTa	Tweets by Spanish politicians
distilbert-base-multilingual-cased-toxicity	102 languages	DistilBERT multilingual	Jigsaw
distilbert-base-german-cased-toxic-comments	de	German DistilBERT	Various incl. GermEval 2018
russian_toxicity_classifier (Dementieva et al., 2022)	ru	RuBERT	Russian Language Toxic Comments
xlmr-large-toxicity-classifier	am, ar, de, en, es, hi, ru, uk, zh	XLM-RoBERTa	TextDetox 2024 (Dementieva et al., 2024b)
amharic-hate-speech	am	Amharic RoBERTa	Amharic Hate Speech
multilingual-toxic-xlm-roberta (Hanu, 2020)	en, es, fr, it, pt, ru, tr	XLM-RoBERTa	Jigsaw Multilingual
toxic-bert (Hanu, 2020)	en	BERT	Jigsaw

Table 2: Open-source toxicity classifiers evaluated in this work.

Model	Type
Llama 3.1 8B Instruct (Grattafiori et al., 2024)	LLM
Gemma 3 4B Instruct (Gemma Team et al., 2025)	LLM
GPT-4o (OpenAI, 2024)	LLM
NLLB 200 3.3B (NLLB-Team et al., 2022)	NMT

Table 3: Translation systems evaluated in this work.

3.2 Datasets

We curate a set of ten toxicity benchmarks for evaluating pipeline performance, spanning 17 languages, where each dataset comprises samples of text with gold labels indicating toxicity. Benchmarks were identified via searching related work on toxicity detection and by searching the Hugging Face datasets catalog. We limited our search to only datasets comprising natural human data, and to those where the gold labels are produced by human annotators, such that datasets comprising model-generated or otherwise synthetic text or labels were discarded. Datasets were restricted to those with a permissive license, where data provenance was clearly indicated, and where the data is readily-accessible online. This resulted in the following benchmarks: Amharic Hate Speech (Ayele et al., 2023); GermEval 2018 (German; Wiegand et al. 2018); Jigsaw Multilingual (Spanish, French, Italian, and Turkish partitions only; Kivlichan et al. 2020); L-HSAB (Levantine Arabic; Mulki et al. 2019); MACD (Hindi, Kannada, Malayalam, Tamil, and Telugu; Gupta et al. 2022); Offensive-Hebrew (Hamad et al., 2023); ToLD-Br (Brazilian Portuguese; Leite et al. 2020); Russian Language Toxic Comments (Belchikov, 2019); Thai Toxicity Tweet Corpus (Sirihattasak et al., 2018); and TextDetox 2024 (Ukrainian partition only; Dementieva et al. 2024a). See Table 1 for full details.

Across all datasets, only the test partition is used for evaluation. Where a toxicity classifier is trained on data that includes one of our benchmark's training partitions, we consider that classifier to be operating ID. Otherwise, as the classifier has been trained on data unlike the benchmark, we consider it to be operating OOD. See Table 2 for the training data used to produce each classifier.

For the purposes of our evaluation, we intentionally avoid drawing a distinction between toxicity detection and and hate speech detection. While hate speech and toxic or offensive are distinct concepts (Davidson et al., 2017; Waseem et al., 2017)—with hate speech typically being interpreted as directed toward a specific group (Davidson et al., 2017; Röttger et al., 2021)—in practice, most evaluation datasets use the terms toxicity, abusive or offensive language, and hate speech almost interchangeably (Fortuna et al., 2020; Banko et al., 2020). As a result, we consider datasets spanning toxicity and hate speech detection, and expect minimal difference in findings between tasks.

3.3 Toxicity classifiers

We consider eight open-source toxicity classifiers, including English-language, non-English monolingual, and multilingual, all of which are available on Hugging Face. See Appendix A.1 for selection

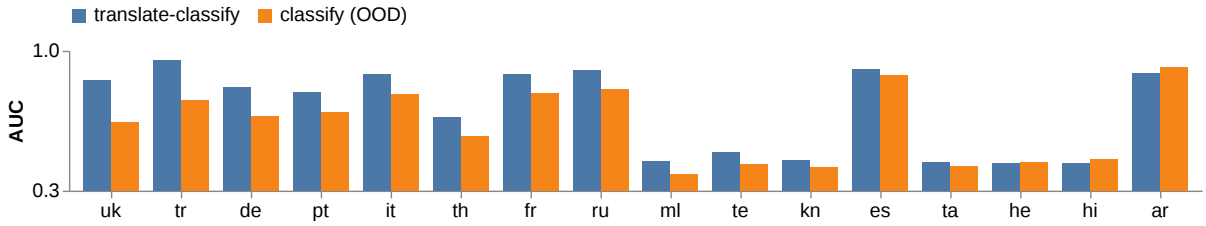


Figure 2: AUC of best possible `translate-classify` pipeline (over all combinations of translation systems and English toxicity classifiers) and best possible `classify (OOD)` pipeline (over all OOD toxicity classifiers). **The `translate-classify` approach wins across 13 out of 16 evaluated languages.**

criteria and Table 2 for full details of all classifiers considered.

All classifiers evaluated make use of pretrained Transformer-based encoder models as a backbone, such as BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2020), ROBERTa (Liu et al., 2019), or the multilingual XLM-ROBERTa (Conneau et al., 2020a), some of which have undergone additional fine-tuning on language specific corpora, such as Russian RuBERT (Kuratov and Arkhipov, 2019). All classifiers are then finetuned on a portion of labeled toxicity data, such as detailed in §3.2.

3.4 Translation systems

For `translate-classify` pipelines, we translate samples into English with a translation system before classifying the translations with an English-supporting classifier. We evaluate four different translation systems (see Table 3), including both encoder-decoder MT systems (NMT) and decoder-only (i.e., LLM) translation systems. In the NMT category, we use NLLB 200 3.3B (NLLB-Team et al., 2022). We evaluate three LLM systems (two open-weights and one behind-API): Llama 3.1 8B Instruct (Grattafiori et al., 2024), Gemma 3 4B Instruct (Gemma Team et al., 2025) and GPT-4o (OpenAI, 2024). The following prompt is used to produce the translations:

```
Translate the following sentence from
    ↪ {{lang}} into English. Respond
    ↪ only with the translation into
    ↪ English, without any additional
    ↪ comments.
{{sentence}}
```

4 Experiments

4.1 Translated pipelines often win

We compare the AUC of the best `translate-classify` pipeline (the best possible combination of translation system and toxicity classifier) against the best possible `classify` pipeline (the best toxicity classifier that supports each language).

Results In Fig. 2, we evaluate `translate-classify` in the common scenario where a language-specific finetuned toxicity classifier is unavailable, i.e., where classifiers are operating OOD with respect to either their source language or training domain, `classify (OOD)`. We observe that in such a scenario, the best `translate-classify` pipeline outperforms the best `classify (OOD)` pipeline across 13 of 16 languages considered (81.3%). Reducing a degree of freedom by using a fixed classifier, `distilbert-base-multilingual-cased-toxicity`, `translate-classify` still outperforms `classify` in 12 of 16 languages (75%; see Fig. S1).

In Fig. 3 we evaluate translated pipelines in scenarios where a language-specific finetuned classifier *is* available (`classify (ID)`), though we note that this is far from the case for the majority of languages. Here, `translate-classify` still offers a robust baseline, outperforming finetuned `classify (ID)` pipelines across three out of seven languages. See Table S1 for full results over all languages.

4.2 Translation benefit scales with resources

Next, we explore which factors determine the success of `translate-classify` pipelines. To allow for consistent comparison across languages and control for variability in classifier performance, we now limit ourselves to two fixed classifiers: for `translate-classify` we use the English classifier, `toxic-bert`, while for `classify` we use our most multilingual classifier, `distilbert-base-multilingual-cased-toxicity`. We evaluate the role of language resourcefulness and translation quality on change in AUC between pipelines, as specified in §3.1.

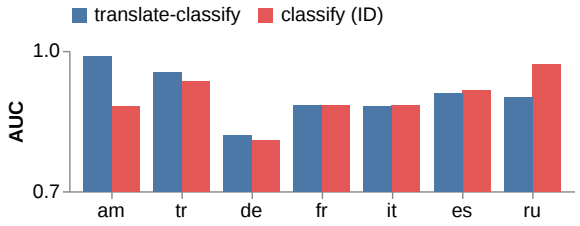


Figure 3: AUC of best possible `translate-classify` pipeline (over all combinations of translation systems and English toxicity classifiers) and best possible `classify (ID)` pipeline (over all ID toxicity classifiers). **The `translate-classify` approach still wins across three of seven languages where in-distribution finetuned classifiers are available.**

Results In Fig. 4, we observe that the relative benefit of `translate-classify` over `classify`, as measured by the change in AUC, is higher for better-resourced languages. This is consistent across four different translation systems, including both LLM and NMT systems. After fixing the best performing classifiers, we notice that the relative benefit of translation for some languages is affected, suggesting the framework is susceptible to model selection to maximize gains.

Similarly, in Fig. 5 we see that the relative benefit of `translate-classify` increases with the quality of translations in each language, across both LLM and NMT systems. We note a higher sensitivity to both language resourcefulness and translation quality for the NMT system, NLLB, compared with LLM systems.

4.3 MT-SFT reduces refusal and improves performance

When using safety-tuned LLMs for translation, a we noticed that key risk is *refusal*: the model declines to translate inputs containing harmful or toxic content, which can severely limit coverage in toxicity detection. We examine whether finetuning for MT can mitigate this problem by comparing two `translate-classify` pipelines: (1) `translate-classify (Llama 3)`, which uses translations from a standard instruction-tuned LLM (Llama 3.1 8B Instruct), and (2) `translate-classify (+TowerBlocks/MT)`, which uses translations from the same base model after supervised finetuning (MT-SFT) on the TowerBlocks/MT dataset (Alves et al., 2024) (see Appendix A.2 for details). Both pipelines feed translations to a fixed English-only classifier, `toxic-bert`, to isolate translation effects, and are compared against a direct multilingual `classify` pipeline using `distilbert-base-multilingual-cased-toxicity`.

Refusal detection We use `Minos` (Suphavadeeprasit et al., 2025) to assign each translation output $y_i = T(x_i)$ a refusal probability $P_r(y_i)$. The refusal rate is defined as:

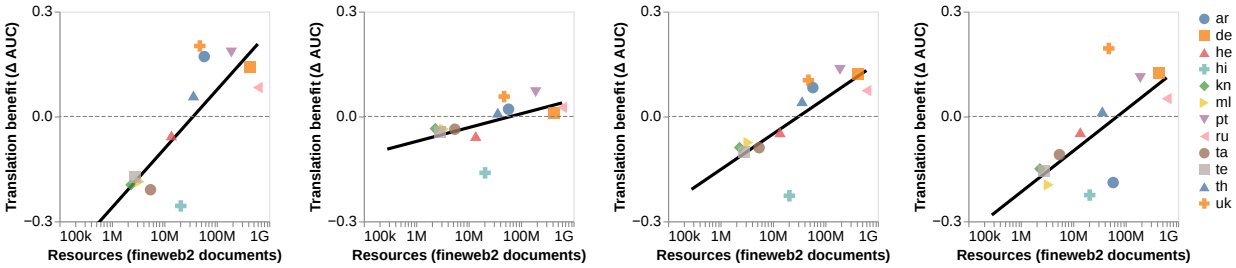
$$\mathrm{R}(T) = \frac{1}{N}\sum_{i=1}^{N}[P_r(T(x_i)) > 0.95],$$

where a 0.95 threshold minimizes false positives. For two systems T_A and T_B, the difference in refusal rates is:

$$\Delta\mathrm{R}(P_{T_A}, P_{T_B}) = \mathrm{R}(P_{T_A}) - \mathrm{R}(P_{T_B}).$$

Refusal results As shown in Fig. 8, `translate-classify (+TowerBlocks/MT)` reduces refusal rates in *every* language compared with `translate-classify (Llama 3)`. The reduction scales approximately log-linearly with language resources (Fig. 9a), indicating that MT-SFT particularly benefits high-resource languages where refusals are rarer but still impactful. Lower refusal means more toxic content is actually processed by the classifier, directly improving pipeline coverage.

Human verification of refusal mitigation To validate both the accuracy of our automated refusal detection and the effectiveness of MT-SFT in addressing refusals, we conducted a targeted human annotation study. For each dataset, we randomly sampled up to 5% of the content flagged as refusals by the base Llama 3.1 8B Instruct model, with a minimum of 10 examples per dataset. Annotators manually verified whether each flagged case was indeed a refusal, then examined translations of the same inputs generated by the MT-finetuned model. As shown in Table 4, the refusal detector achieved perfect true positive rates for Thai, German, and Ukrainian, and high —though not perfect— accuracy for Malayalam and Levantine Arabic, where some false positives were observed. Importantly, the MT-finetuned model produced valid translations for *all* annotated examples, yielding a true negative rate of 100% across every language in the sample. This confirms that, at least for the languages tested, MT-SFT can completely eliminate refusals observed in the base instruction-tuned model, turning previously blocked content into usable inputs for the downstream classifier.



(a) GPT-4o (b) Llama 3.1 8B Instruct (c) Gemma 3 4B Instruct (d) NLLB

Figure 4: Change in AUC (i.e., translation benefit) between `translate-classify` pipelines with a fixed English classifier, `toxic-bert`, and `classify` pipelines with a fixed multilingual classifier, `distilbert-base-multilingual-cased-toxicity`, as a function of language resources, over four translation systems **(a)** GPT-4o, **(b)** Llama 3.1 8B Instruct, **(c)** Gemma 3 4B Instruct, and **(d)** NLLB. **Translation benefit is increased for higher resourced languages.**

(a) Llama 3.1 8B Instruct (b) NLLB

Figure 5: Change in AUC (i.e., translation benefit) between `translate-classify` pipelines and `classify` pipelines, as a function of English translation quality measured by `CometKiwi-DA-XL`, over two translation systems **(a)** Llama 3.1 8B Instruct, and **(b)** NLLB. **Translation benefit increases with translation quality for both LLM-based and NMT systems.**

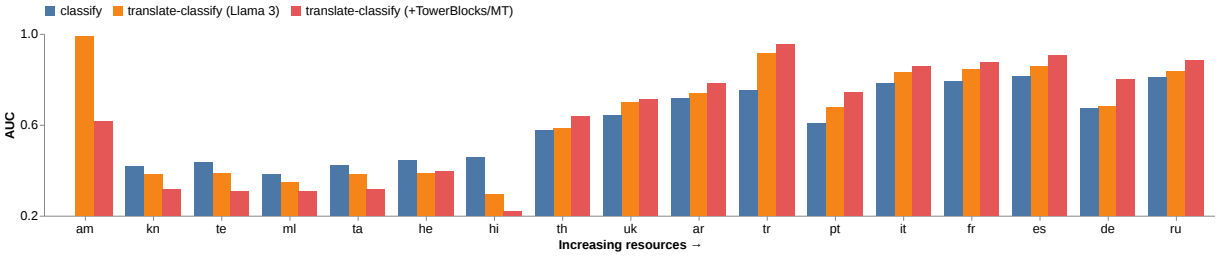


Figure 6: AUC of `translate-classify (Llama 3)` and `translate-classify (+TowerBlocks/MT)` using a fixed English classifier, `toxic-bert`, and a `classify` pipeline using a fixed multilingual classifier, `distilbert-base-multilingual-cased-toxicity`. **Using a finetuned LLM for translation improves pipeline performance for higher-resourced languages.**

MT-SFT improves performance for high resource languages In addition to lowering refusals, MT-SFT also improves classification accuracy. In Fig. 6, `translate-classify (+TowerBlocks/MT)` achieves higher AUC than `translate-classify (Llama 3)` for 11 of 17 languages, with gains concentrated in high-resource settings. When measured against the multilingual `classify` baseline, `translate-classify (+TowerBlocks/MT)` shows even stronger sensitivity to language resource availability (Fig. 7).

4.4 LLM judges underperform on lower-resourced languages

Given the strong performance of LLMs across a range of tasks, we additionally compare pipelines based on traditional classifiers vs. zero-shot LLM judges.

Specifically, we analyze the performance of

Name	TPR (`Llama 3.1`)	TNR (`+TowerBlocks/MT`)
TH	100%	100%
DE	100%	100%
UK	70%	100%
ML	75%	100%
AR	40%	100%

Table 4: Analysis of human annotations of refusal predictions, showing True Positive Rate (TPR) of Llama 3.1 8B Instruct (`Llama 3.1`) and the True Negative Rate (TNR) of the same model finetuned on TowerBlocks/MT (`+TowerBlocks/MT`). **Refusal detection is highly accurate for Thai, German, and Ukrainian. Even for Malayalam and Levantine Arabic where the refusal detector produces some false positives, the finetuned model never refuses.**

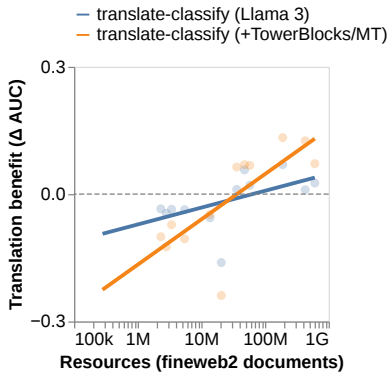


Figure 7: Change in AUC (i.e., translation benefit) of `translate-classify (Llama 3)` and `translate-classify (+TowerBlocks/MT)` pipelines with a fixed English classifier against a fixed `classify` pipeline with a fixed multilingual classifier, as a function of language resources. **The `translate-classify (+TowerBlocks/MT)` pipeline performance is more sensitive to available language resources.**

two judge-based pipelines: (1) `judge`, where an LLM is prompted to classify untranslated samples, and (2) `translate-judge`, where the LLM is prompted to classify samples that have been translated into English. The judge in both cases is Llama 3.1 8B Instruct with the following custom prompt, where `{{sample}}` is either the translated or original sample according to pipeline:

```
You are a toxicity classifier. Classify
    ↪ the following text as toxic or
    ↪ not toxic. Respond with 'toxic'
    ↪ or 'not toxic' with no other
    ↪ comments or explanation.
Text: {{sample}}
Classification:
```

Translations are also produced using Llama 3.1 8B Instruct. We compare both judge pipelines against a `classify` pipeline based on the overall best English classifier, `toxic-bert`, coupled again with Llama 3.1 8B Instruct for fair comparison. We evaluate all pipelines using accuracy as opposed to AUC, as the LLM judge emits only a label, rather than a continuous score.

Results Figure 10 illustrates notable patterns in the comparative performance of `translate-judge` and `judge` pipelines. Across all languages, translation-based approaches narrowly outperform their untranslated counterparts; however, this advantage becomes pronounced in low-resource settings, where `translate-judge` completely dominates, outperforming `judge` in 6 out of 7 low-resource languages. Similarly, `translate-classify` pipelines provide a slight overall edge compared to both `judge` and `translate-judge`, but the margin is especially significant for low-resource languages, where `translate-classify` overwhelmingly wins (again in 6 out of 7 cases). These results further indicate that multilingual capabilities in LLMs are not homogeneously distributed: while MT models demonstrate broader multilingual reach, toxicity classification performance by LLMs is markedly less consistent across lower-resource languages.

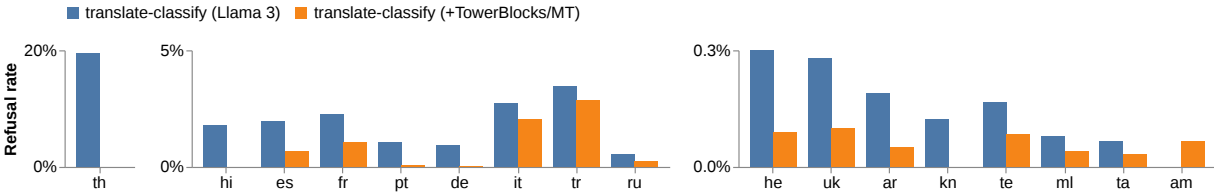
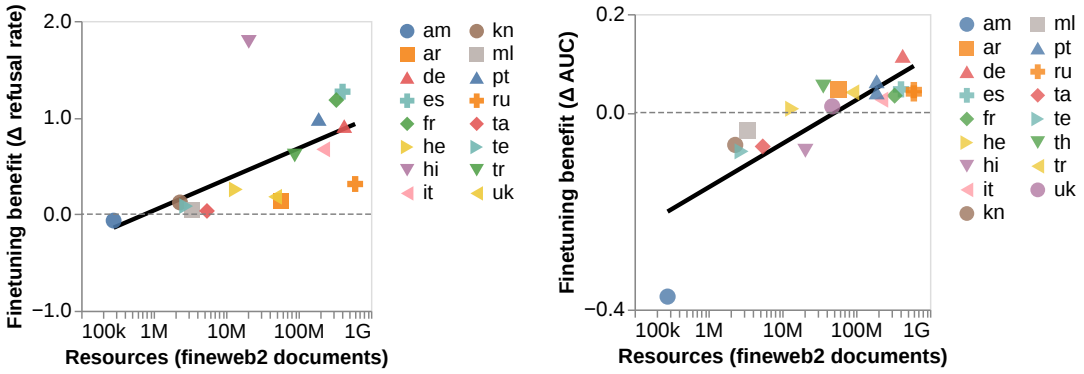


Figure 8: Translation refusal rate of `translate-classify (Llama 3)` and `translate-classify (+TowerBlocks/MT)` pipelines. Note three separate scales for legibility. **Using a finetuned LLM for translation reduces refusal rates.**



(a) Δ refusal

(b) Δ AUC

Figure 9: Change in **(a)** translation refusal rate and **(b)** AUC of a `translate-classify (+TowerBlocks/MT)` pipeline against a `translate-classify (Llama 3)` pipeline, both with a fixed English classifier, `toxic-bert`, as a function of language resources. **The benefit of using a finetuned LLM for translation, in terms of both refusal rates and improved performance, increases for with language resources.**

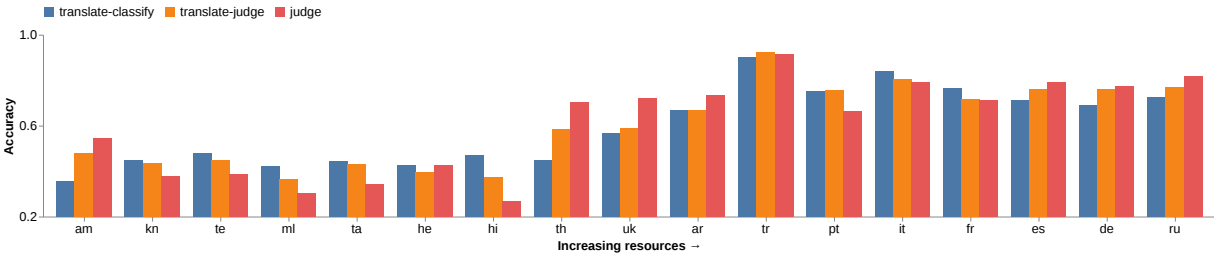


Figure 10: Accuracy of `translate-judge`, `judge`, and `translate-classify` with a fixed English classifier, `toxic-bert`, all using a Llama 3 for translation. **Translation with traditional classifiers outperforms LLM judges for most lower resourced languages.**

5 Discussion

Across ten benchmarks spanning 17 languages, our analysis suggests that translation-based approaches can be successfully leveraged to support multilingual toxicity detection at scale. Specifically, we observe that `translate-classify` pipelines outperform `classify (OOD)`, a non-finetuned classifier operating OOD (i.e., an off-the-shelf model) in the majority of cases, and can even occasionally outperform `classify (ID)`, dedicated finetuned classifiers evaluated ID. The relative benefit of using `translate-classify` over `classify` pipelines increases with both a language's available resources and the quality of the translation system. This may suggest that while translation may be an effective strategy *in general*, it does have the potential to increase performance disparities between better- and worse-resourced languages. We additionally note that using an MT-finetuned LLM for translations can further drive up pipeline performance, in part, by reducing refusal rates, but that this benefit appears to be reserved for higher-resourced languages. Finally, we evaluate the utility of an LLM judge approach over traditional (e.g., BERT-based) classification, finding that in lower-resourced languages,

`translate-classify` consistently outperforms.

Practical recommendations We make four practical recommendations for practitioners looking to deploy multilingual toxicity detection at scale.

1. At the very least, `translate-classify` pipelines using traditional classifiers and LLM-based translation should be considered a robust baseline.

2. If fine-tuning on dedicated data is unavailable, a `translate-classify` pipeline is likely to provide a strong first choice of model, particularly in languages where translation quality is high.

3. If operating on a higher-resourced language, making use of an MT-finetuned LLM may offer some performance improvements over a standard instruction-tuned LLM, particularly in the scenario where refusal rates can be reduced.

4. Unlike many other NLP tasks, an LLM judge demonstrates only a limited performance advantage on select higher-resourced languages when compared to traditional (e.g., BERT-based) classifiers.

Limitations

While we approach multilingual toxicity detection through the lens of a practitioner making a choice between available, off-the-shelf pipeline components, this does limit our ability to analyze the role of specific finetuning details. For example, in contrast with previous work (Artetxe et al., 2023) that has contrasted cross-lingual transfer pipelines where the classifier was finetuned on either the original domain or the outputs of the translation system, we only make use of publicly-available classifiers which may be finetuned on different numbers of samples or different domains, and none of which are finetuned on translations. However, given the performance improvements offered by the `translate-classify` pipeline *without finetuning on translations*, we might expect a translation-finetuned classifier to further benefit the `translate-classify` approach.

As we note in §3.2, our work is also potentially limited by shifts in data distribution between languages. In order to identify broad trends across many languages with different levels of resources, we draw samples from different constituent datasets. These datasets, however, are drawn from different domains (e.g., social media vs. WikiMedia talk pages) with labels produced using different annotation schemas (e.g., identifying hate speech vs. toxicity). As a result, our conclusions should be interpreted as indicative of general trends about the relative utility of translation, rather than individual claims about how well translation may function on any given language. This limitation could be overcome with access to additional highly-multilingual datasets of labeled toxicity data.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Waleed Ammar, Phoebe Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 152–157, Berlin, Germany. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of EMNLP*, pages 7674–7684.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. In *Transactions of the Association for Computational Linguistics*, volume 7, pages 597–610.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023. Exploring Amharic Hate Speech Data Collection and Classification Approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In

Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 125–137, Online. Association for Computational Linguistics.

Anatoly Belchikov. 2019. Russian Language Toxic Comments.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.

Marta Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9570–9586, Singapore. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025, Abu Dhabi, UAE. Association for Computational Linguistics.

Daryna Dementieva, Valeriia Khylenko, Nikolay Babakov, and Georg Groh. 2024a. Toxicity Classification in Ukrainian. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 244–255, Mexico City, Mexico. Association for Computational Linguistics.

Daryna Dementieva, Varvara Logacheva, Irina Nikishina, Alena Fenogenova, David Dale, Irina Krotova, Nikita Semenov, Tatiana Shavrina, and Alexander Panchenko. 2022. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*, pages 114–131. RSUH.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024b. Overview of the Multilingual Text Detoxification Task at PAN 2024. *CLEF 2024: Conference and Labs of the Evaluation Forum*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, pages 18–24.

Anni Eskelinen, Laura Silvala, Filip Ginter, Sampo Pyysalo, and Veronika Laippala. 2023. Toxicity detection in Finnish using machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 685–697, Tórshavn, Faroe Islands. University of Tartu Library.

Julen Etxaniz, Mikel Artetxe, and Rodrigo Agerri. 2023a. On the effectiveness of translate-test for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9933–9947, Singapore. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023b. Do multilingual language models think better in english? *Preprint*, arXiv:2308.01223.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, Gothenburg, Sweden. Association for Computational Linguistics.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Gemma Team, Aishwarya Kamath, Johan Ferret, ..., Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Alfio Gliozzo and Carlo Strapparava. 2006. Exploiting lexical alignment for cross-language textual entailment. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–40, Trento, Italy. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, ..., Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Vikram Gupta. 2021. Multilingual and multilabel emotion recognition using virtual adversarial training. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 74–85, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, Hastagiri Prakash Vanchinathan, and Animesh Mukherjee. 2022. Multilingual Abusive Comment Detection at Scale for Indic Languages. *Advances in Neural Information Processing Systems*, 35:26176–26191.

Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, and Nadim Nashif. 2023. Offensive Hebrew Corpus and Detection using BERT. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8.

Laura Hanu. 2020. Detoxify.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw Multilingual Toxic Comment Classification.

Jordan K. Kobellarz and Thiago H. Silva. 2022. Should we translate? evaluating toxicity in online comments when translating from portuguese to english. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '22, page 89–98, New York, NY, USA. Association for Computing Machinery.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language. *Preprint*, arXiv:1905.07213.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. In *arXiv preprint arXiv:1309.4168*.

Niklas Muennighoff, Nouamane Tazi, and Sebastian Ruder. 2022. Crosslingual generalization through multitask finetuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1596–1610.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Omnilingual MT Team, Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET: Dataset, Benchmark and Open initiative for Universal Quality Evaluation in Translation. *Preprint*, arXiv:2502.04314.

OpenAI. 2024. GPT-4o System Card. *Preprint*, arXiv:2410.21276.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein

Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *Preprint*, arXiv:2506.20920.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of EMNLP*, pages 7654–7673.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer. *Preprint*, arXiv:2107.11353.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 320–322, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. *Preprint*, arXiv:2309.11925.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Sugan Sirihattasak, Mamoru Komachi, and H. Ishikawa. 2018. Annotation and Classification of Toxicity for Thai Twitter. In *Proceedings of TA-COS 2018 – 2nd Workshop on Text Analytics for Cybersecurity and Online Safety*.

Jai Suphavadeeprasit, Teknium, Chen Guang, Shannon Sands, and rparikh007. 2025. Minos classifier.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing*, pages 235–243, Suntec, Singapore. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. Technical report, Verlag der Österreichischen Akademie der Wissenschaften.

A Additional methods

A.1 Toxicity classifier selection

We evaluate on a sample of toxicity classifiers that are publicly-available on Hugging Face. We reviewed classifiers that matched the search terms "toxic" and "toxicity", selecting those that supported either English or one or more of the 17 languages analyzed. Classifiers were limited to those that were permissively-licensed, with clear data provenance (to allow for distinguishing between ID and OOD performance), and substantial community engagement (as measured by downloads and likes). See Table 2 for all classifiers evaluated.

A.2 MT finetuning an LLM

We used Llama 3.1 8b Instruct as our baseline model and finetuned it for 5 epochs with the MT split from Towerblocks 0.2, a multi-task, multilingual SFT dataset. We employed the AdamW optimizer with a learning rate initialized to 1×10^{-6}, β_1 and β_2 coefficients set to 0.9 and 0.95 respectively, and a weight decay of 0.1. We used a cosine annealing learning rate scheduler configured with a final learning rate scaled to 0.2 times the initial rate and a total of 1,000 warmup steps.

B Additional results

In Table S1 we present the detailed results behind Figs. 2 and 3, showing the performance of the best-possible `translate-classify`, `classify (OOD)`, and `classify (ID)` pipelines over all languages. In Tables S2 to S4 we present the corresponding best-performing translation system and classifier combinations for `translate-classify`, `classify (OOD)`, and `classify (ID)` respectively.

In Fig. S1, we present a version of Fig. 2 but reducing one degree of freedom: rather than choosing the best-possible combination of translation system and classifier, here we choose the best possible translation system though use a fixed classifier, `distilbert-base-multilingual-cased-toxicity`. In this setting, `translate-classify` still outperforms across 12 of 16 languages.

	AUC		
Language	ID	OOD	Translated
ar	-	**0.92**	0.89
he	-	**0.44**	0.44
hi	-	**0.46**	0.44
kn	-	0.42	**0.45**
ml	-	0.38	**0.45**
pt	-	0.69	**0.79**
ta	-	0.42	**0.44**
te	-	0.43	**0.49**
th	-	0.57	**0.67**
uk	-	0.64	**0.85**
am	0.88	-	**0.99**
de	0.81	0.67	**0.82**
es	**0.92**	0.88	0.91
fr	**0.88**	0.79	0.88
it	**0.88**	0.78	0.88
ru	**0.97**	0.81	0.90
tr	0.94	0.75	**0.96**

Table S1: Best possible performance over all languages. Where a finetuned classifier isn't available, translation-based pipelines often outperform.

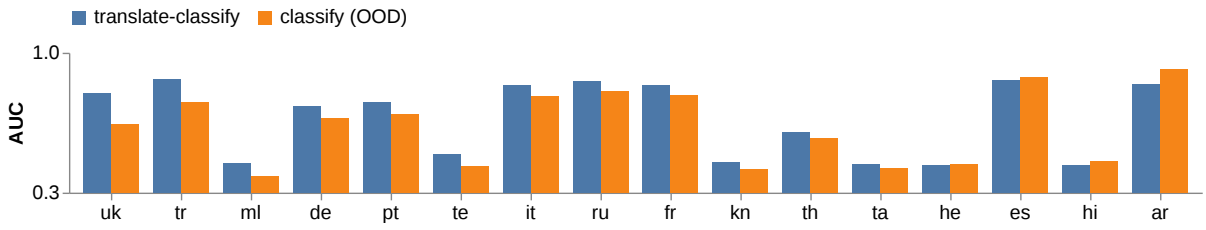


Fig. S1: Translation-based toxicity detection pipelines with a fixed English-supporting classifier, `distilbert-base-multilingual-cased-toxicity`, outperform off-the-shelf pipelines across 12 out of 16 evaluated languages.

Language	Classifier	AUC
am	textdetox/xlmr-large-toxicity-classifier	0.88
de	ml6team/distilbert-base-german-cased-toxic-comments	0.81
es	unitary/multilingual-toxic-xlm-roberta	0.92
fr	unitary/multilingual-toxic-xlm-roberta	0.88
it	unitary/multilingual-toxic-xlm-roberta	0.88
ru	s-nlp/russian_toxicity_classifier	0.97
tr	unitary/multilingual-toxic-xlm-roberta	0.94

Table S2: Best-performing ID pipeline per language.

Language	Classifier	AUC
ar	textdetox/xlmr-large-toxicity-classifier	0.92
de	citizenlab/distilbert-base-multilingual-cased-toxicity	0.67
es	textdetox/xlmr-large-toxicity-classifier	0.88
fr	citizenlab/distilbert-base-multilingual-cased-toxicity	0.79
he	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
hi	citizenlab/distilbert-base-multilingual-cased-toxicity	0.46
it	citizenlab/distilbert-base-multilingual-cased-toxicity	0.78
kn	citizenlab/distilbert-base-multilingual-cased-toxicity	0.42
ml	citizenlab/distilbert-base-multilingual-cased-toxicity	0.38
pt	unitary/multilingual-toxic-xlm-roberta	0.69
ru	citizenlab/distilbert-base-multilingual-cased-toxicity	0.81
ta	citizenlab/distilbert-base-multilingual-cased-toxicity	0.42
te	citizenlab/distilbert-base-multilingual-cased-toxicity	0.43
th	citizenlab/distilbert-base-multilingual-cased-toxicity	0.57
tr	citizenlab/distilbert-base-multilingual-cased-toxicity	0.75
uk	citizenlab/distilbert-base-multilingual-cased-toxicity	0.64

Table S3: Best-performing OOD pipeline per language.

Language	Translation system	Classifier	AUC
am	Llama 3.1 8B Instruct	unitary/toxic-bert	0.99
ar	GPT-4o	unitary/toxic-bert	0.89
de	GPT-4o	unitary/multilingual-toxic-xlm-roberta	0.82
es	GPT-4o	unitary/toxic-bert	0.91
fr	GPT-4o	unitary/toxic-bert	0.88
he	Llama 3.1 8B TowerBlocks	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
hi	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
it	GPT-4o	unitary/toxic-bert	0.88
kn	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.45
ml	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.45
pt	GPT-4o	unitary/toxic-bert	0.79
ru	GPT-4o	unitary/multilingual-toxic-xlm-roberta	0.90
ta	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.44
te	Llama 3.1 8B Instruct	citizenlab/distilbert-base-multilingual-cased-toxicity	0.49
th	GPT-4o	textdetox/xlmr-large-toxicity-classifier	0.67
tr	Llama 3.1 8B TowerBlocks	unitary/toxic-bert	0.96
uk	GPT-4o	unitary/multilingual-toxic-xlm-roberta	0.85

Table S4: Best-performing translated pipeline per language.

Feeding Two Birds or Favoring One? Adequacy–Fluency Tradeoffs in Evaluation and Meta-Evaluation of Machine Translation

Behzad Shayegh[1,*] **Jan-Thorsten Peter**[2] **David Vilar**[2] **Tobias Domhan**[2]
Juraj Juraska[2] **Markus Freitag**[2] **Lili Mou**[1,3]
[1]Dept. Computing Science, Alberta Machine Intelligence Institute (Amii), University of Alberta
[2]Google [3]Canada CIFAR AI Chair, Amii
{the.shayegh, doublepower.mou}@gmail.com
{vilar, jtp, domhant, jjuraska, freitag}@google.com

Abstract

We investigate the tradeoff between adequacy and fluency in machine translation. We show the severity of this tradeoff at the evaluation level and analyze where popular metrics fall within it. Essentially, current metrics generally lean toward adequacy, meaning that their scores correlate more strongly with the adequacy of translations than with fluency. More importantly, we find that this tradeoff also persists at the meta-evaluation level, and that the standard WMT meta-evaluation favors adequacy-oriented metrics over fluency-oriented ones. We show that this bias is partially attributed to the composition of the systems included in the meta-evaluation datasets. To control this bias, we propose a method that synthesizes translation systems in meta-evaluation. Our findings highlight the importance of understanding this tradeoff in meta-evaluation and its impact on metric rankings.

1 Introduction

As translation systems become more sophisticated and widely adopted, the critical challenge of accurately evaluating their performance has come to the forefront, driving significant ongoing research to the developement of more accurate and robust metrics (Chatzikoumi, 2020; Guerreiro et al., 2024). Traditionally, translation evaluation relied on lexical metrics such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), which primarily consider the n-gram overlap between the candidate and reference translations. However, these metrics have been shown insufficient for measuring the quality of modern translation systems (Babych and Hartley, 2008; Freitag et al., 2022). In the recent decade, researchers have been exploring the idea of training neural models to measure translation quality (Wieting et al., 2019; Ma et al., 2019; Freitag et al., 2022; Guerreiro et al., 2024), with popular examples including MetricX (Juraska et al., 2023, 2024) and Comet (Rei et al., 2020, 2022a).

Two key aspects of translation quality, long discussed in the community (Pierce and Carroll, 1966; White and O'Connell, 1993; Banchs et al., 2015; Martindale and Carpuat, 2018), are[1]:

- *Fluency*: the grammatical correctness and naturalness of the translation; and
- *Adequacy*: how well the translation conveys the meaning of the source text.

Flamich et al. (2025) demonstrate that an adequacy–fluency tradeoff exists in translation: optimizing for one aspect eventually sacrifices the other. They further introduce measurements to study the severity of this tradeoff at **the level of translation systems**.

The aforementioned tradeoff naturally results in an adequacy–fluency tradeoff at **the level of evaluation metrics**, if we limit our evaluation to a single score: increasing the capability of measuring one aspect eventually comes at the cost of decreasing the capability of measuring the other. It is important to understand this tradeoff and know the position of each metric to avoid undesired biases when optimizing translation systems according to the metric.

In this work, we demonstrate the severity of the adequacy–fluency tradeoff at the evaluation level. We analyze current evaluation setups in WMT (Callison-Burch et al., 2007; Moghe et al., 2025), and show that there is a significant disagreement between adequacy and fluency when ranking systems (Table 1). This directly imposes a severe tradeoff, as a metric can only agree with either adequacy or fluency when comparing discordant system pairs. Subsequently, we empirically analyze several contemporary translation metrics to illustrate their positions within this tradeoff.

More importantly, we show that this adequacy–

[*]Work partially done when the first author was interning at Google.

[1]Different namings and slightly different definitions are used in the literature for these two aspects; however, the main idea remains consistent.

Proceedings of the Tenth Conference on Machine Translation, pages 269–285
November 8-9, 2025

Evaluation Set	Concordance	Discordance
En–De'23	49 (74%)	17 (26%)
Zh–En'23	70 (67%)	35 (33%)
En–De'24	98 (72%)	38 (28%)
En–Es'24	55 (71%)	23 (29%)
Ja–Zh'24	58 (74%)	20 (26%)

Table 1: Concordance and discordance between adequacy and fluency in MQM datasets (§3.1), reported as system pair counts and percentages. Concordance means one system in a pair outperforms the other in both metrics; discordance indicates inconsistent performance across metrics. Notice that 26–33% discordance, although being the minority, is way far from being negligible and imposes a significant tradeoff when the metric quality improves.

fluency tradeoff extends even to **the level of meta-evaluation** (the evaluation of evaluation metrics). Meta-evaluation typically compares the scores of a set of candidates given by a metric to those given by humans (Callison-Burch et al., 2007). We show that the selection of the candidates used during the meta-evaluation significantly influences the identified optimal metric within the tradeoff: if there is a higher variance in the candidates' adequacy than in their fluency, the meta-evaluation will lean toward favoring metrics that prioritize adequacy. Conversely, if the fluency of these candidates shows higher variance, the meta-evaluation will prefer fluency-oriented metrics. This can inadvertently guide the development of evaluation metrics, and thus the development of translation systems, toward a particular bias toward either adequacy or fluency. We propose to synthesize candidates with desired variance in their adequacy and fluency scores to conduct a controlled (balanced) meta-evaluation.

2 Background & Related Work

2.1 Translation Metrics

The evaluation of translation systems is a critical aspect of their development and deployment, and has been researched for years (Tao et al., 2018; Chatzikoumi, 2020). This research has resulted in the development of a variety of metrics (Chauhan and Daniel, 2023; Guerreiro et al., 2024).

Traditional metrics, such as BLEU (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), and ChrF (Popović, 2015), compare the output of a translation system against one or more human-created reference translations, only at the surface level (e.g., based on string overlap). While widely adopted for their simplicity, these surface-level metrics are shown to be incapable of measuring the quality of modern, high-quality translation systems (Freitag et al., 2022).

Recent advancements have led to the development of *trained metrics*, which leverage machine learning models to assess translation quality (Wieting et al., 2019; Freitag et al., 2022). These metrics, such as MetricX (Juraska et al., 2023, 2024) and Comet (Rei et al., 2020, 2022a), are often trained on large datasets of human annotations, allowing them to learn more nuanced correlations between system outputs and perceived quality.

Translation metrics can be divided into two categories: *reference-based* and *reference-free*. Reference-based metrics, such as those mentioned previously, rely on access to a trusted reference translation and compare the candidate against the reference to assess quality. Reference-free metrics, also known as Quality Estimation (QE) metrics, directly evaluate the candidate given the source text, without requiring a reference translation (Ito et al., 2025). This capability is particularly valuable in scenarios where human references are scarce or impractical to obtain, offering a more flexible and potentially more accurate evaluation paradigm for contemporary translation systems (Agrawal et al., 2021). MetricX and Comet offer their reference-free versions: MetricX QE (Juraska et al., 2023, 2024) and CometKiwi (Rei et al., 2022b).

In our work, we empirically analyze several contemporary translation metrics to illustrate their positions within the adequacy–fluency tradeoff, providing insights into their strengths and limitations.

2.2 Meta-Evaluation

Meta-evaluation assesses how well a translation evaluation metric correlates with human annotations (Callison-Burch et al., 2007; Macháček and Bojar, 2013). This is traditionally accomplished by computing Pearson, Kendall, and Spearman correlation coefficients (Macháček and Bojar, 2013; Mathur et al., 2020; Freitag et al., 2023), at either the segment[2] level (for individual translation scores) or the system level (for scores averaged per system). More recently, pairwise accuracy (PA; Kocmi et al., 2021) and soft pairwise accuracy (SPA; Thompson et al., 2024) have gained prominence as they directly assess how well a met-

[2]Following the common terminology in the community, we refer to each data sample as a *segment*.

ric aligns with human preferences over pairs of translations (Mathur et al., 2020). Given their popularity, we utilize these two metrics for our analysis and elaborate their formulations below.

Pairwise accuracy. PA assesses whether a metric correctly identifies the better translation (segment level) or the better system (system level) given a pair. Let m and h be the metric and human scores, respectively. PA is formulated as

$$\frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} \mathbb{1}[\text{sgn}(m_j - m_i) = \text{sgn}(h_j - h_i)] \quad (1)$$

where $\mathcal{P}$ is the set of pairs, and sgn is the sign function. Pairs with tied human scores are excluded (Kocmi et al., 2021).

A key challenge for PA at the segment level is to handle tied metric scores. While techniques like tie calibration are employed to address this issue (Deutsch et al., 2023), recent studies (Perrella et al., 2024) indicate that such approaches are not reliable and the matter warrants further investigation. This challenge is minor at the system level, since averaging scores over many translations rarely yields ties. We therefore focus on system-level PA in our work.

Soft pairwise accuracy. SPA is a system-level meta-metric that extends PA by incorporating statistical significance from both human and metric scores. Unlike PA's binary agreement, SPA compares p-values from hypothesis tests to determine if one system is statistically superior. The SPA score is formulated as

$$\frac{1}{|\mathcal{P}|} \sum_{i,j \in \mathcal{P}} (1 - |p(h_j > h_i) - p(m_j > m_i)|) \quad (2)$$

Here, $p(x_j > x_i)$ is the p-value from a permutation test (Welch, 1990) assessing if system j is superior to system i based on scores x. SPA is sensitive to the magnitude of the score differences, reflecting the metric's confidence in its preference, and compares this confidence against that of human judgments.

3 Experimental Setup

3.1 Dataset

The Multidimensional Quality Metrics (MQM) framework provides a standardized and granular guideline for characterizing and classifying errors in translations (Mariana et al., 2015). Instead of being a score-based annotation, the MQM framework annotates translation errors. Each error is annotated as a substring of the translation with a category (such as accuracy, fluency, terminology, and style) and a severity level (usually major, minor, and neutral). This framework enables a diagnostic understanding of translation system performance, as well as more reliable and standardized evaluation scores.

WMT adopts MQM annotations (Freitag et al., 2021a) for human evaluation, producing datasets of detailed translation error annotations. We use this dataset of years 2023 and 2024 in our meta-evaluation and analysis. We reserve data from prior years (2020–2022) for training to avoid leakage. Our setup is consistent with practice in previous studies (Juraska et al., 2024; Rei et al., 2022a) for fair comparison.

It is common to transform MQM annotations into a single score (Freitag et al., 2022, 2023). This involves assigning a specific penalty to each error based on its category and severity (Freitag et al., 2021b). These penalties are then summed across all errors in the translation. We refer to this aggregated value as the *All MQM* score, where a lower score indicates higher quality.

For a more granular analysis, Flamich et al. (2025) categorize MQM errors into adequacy and fluency classes (Tables 8 and 9 in Appendix A), resulting in *Adequacy MQM* and *Fluency MQM* scores for each segment. We use these specialized scores for our tradeoff analysis. We present detailed statistics for our dataset in Appendix B.

3.2 Meta-Metrics

To study the performance of evaluation metrics in the adequacy–fluency tradeoff, we separately assess their ability to measure translation adequacy and fluency. To achieve this, we compute meta-metrics—namely, PA and SPA (§2.2)—with respect to Adequacy MQM and Fluency MQM. In Appendix C, we point out the necessity of including both PA and SPA in such studies due to their potentially different behavior in measuring bias toward adequacy or fluency.

3.3 Translation Metrics in Our Analysis

In our work, we select a set of diverse and widely used metrics to analyze their positions within the adequacy–fluency tradeoff. These metrics include **BLEU** (Papineni et al., 2002) and **ChrF** (Popović,

2015), representing overlap-based metrics; **MetricX** (Juraska et al., 2024) and **Comet** (Rei et al., 2022a), representing trained metrics; and **MetricX QE** and **CometKiwi** (Rei et al., 2023), representing reference-free metrics.

Moreover, we introduce a set of extreme translation metrics designed to measure only adequacy or fluency, which will provide a clearer view of the tradeoff between the two. Specifically, we develop **AdequacyX** and its reference-free variant, **AdequacyX QE**, trained exclusively on MQM errors categorized under adequacy. Likewise, we develop **FluencyX** trained on fluency annotations.

Technically, these trainable metrics adopt the same neural architecture as MetricX (Juraska et al., 2024), and are fine-tuned from an mT5 checkpoint (Xue et al., 2021). Following Juraska et al. (2024), we train these models on MQM annotations (WMT 2022), covering En–De, En–Ru, and En–Zh.[3] AdequacyX and AdequacyX QE are trained to predict Adequacy MQM, while FluencyX is trained to predict Fluency MQM. Unlike Juraska et al. (2024), we exclusively use MQM annotations for fine-tuning, omitting direct assessments (DA) and synthetic data, as these do not offer a clear adequacy–fluency distinction.

It should be noted that AdequacyX QE does not take the reference translation as input as it is a reference-free metric. We further restrict FluencyX to the candidate translation alone (no source or reference), similar to $\text{SENTINEL}_{\text{CAND}}$ in Perrella et al. (2024), given that fluency is considered source-independent.

To have a fair comparison, we also train our own MetricX and MetricX QE variants; they are identical to AdequacyX (QE), except that they predict All MQM scores. The key difference between our variants and the original MetricX (QE) is the exclusion of DA and synthetic data.

Additionally, we include the log-perplexities of a pretrained Gemma model as a fluency measure, following Flamich et al. (2025). We use **Gemma 3 4B** (Gemma Team, 2025) in this work.

4 Analysis: Tradeoff at the Meta-Evaluation Level

In this section, we explore the adequacy–fluency tradeoff at the meta-evaluation level. We first discuss in §4.1 how the tradeoff exists in meta-evaluation, and show that the imbalance in WMT meta-evaluation datasets imposes a bias that favors adequacy-oriented metrics. We argue that this bias consists of both an intrinsic and an extrinsic component, and that we aim to control the latter. In §4.2, we design a metric to quantify the extrinsic bias, and in §4.3, we propose a practical approach to reduce it. Finally we demonstrate in §4.4 the impact of meta-evaluation bias on the ranking of translation metrics by comparing this ranking before and after reducing the meta-evaluation extrinsic bias.

4.1 Understanding the Meta-Evaluation Bias

Translation meta-evaluation compares the ranking of translation systems produced by a given metric against the ranking based on human annotations, which in our case is the All MQM score. In this section, we will show that the All MQM ranking reflects systems' adequacy more than their fluency, leading the WMT meta-evaluation to favor adequacy-oriented metrics.

For simplicity, we assume that All MQM = Adequacy MQM+Fluency MQM.[4] Therefore, the ranking by All MQM (thus the meta-evaluation assessment) is more influenced by the component with higher variance, and we say the meta-evaluation is *biased* toward that component.

As shown in Table 2, Adequacy MQM exhibits a much higher variance than Fluency MQM in system level scores, consequently having a greater influence on the All MQM system ranking and the meta-evaluation system-level assessment. This influence is further demonstrated by the stronger alignment between All MQM and Adequacy MQM in terms of both PA and SPA. The question now is: *Should we embrace this dominance of adequacy, or is it a bias we should mitigate?*

Answering the above question requires noticing that the higher variance of the system-level Adequacy MQM scores, and consequently its greater influence, can be attributed to two possible causes:

- *Intrinsic variation,* which is due to the preference of the MQM framework and annotators. For example, adequacy errors may be generally considered more severe, leading to larger assigned penalties and thus greater variance.
- *Extrinsic variation,* which is due to the choice of translation systems during meta-evaluation. For example, we may select systems that

[3] We limit En–Zh to the conversation, e-commerce, and social domains, in line with Juraska et al. (2024).

[4] Although there are some errors that are not categorized as either adequacy or fluency, they are rare and have minor influence; therefore, we omit them for simplicity.

	Adequacy MQM				Fluency MQM				
	Variance	F-stat	PA	SPA	Variance	F-stat	PA	SPA	$B(\Delta p)$
En–De'23	0.31	36.5	0.98	0.97	0.06	7.0	0.76	0.75	0.08^A
Zh–En'23	0.23	80.6	0.97	0.93	0.03	12.9	0.70	0.74	0.03^A
En–De'24	0.24	13.7	0.88	0.87	0.11	9.5	0.85	0.81	0.04^A
En–Es'24	0.61	26.4	0.96	0.94	0.28	4.6	0.74	0.77	0.13^A
Ja–Zh'24	0.40	35.3	1.00	0.98	0.10	4.8	0.74	0.75	0.12^A
Macro-Avg	0.36	38.5	0.96	0.94	0.12	7.7	0.76	0.76	

Table 2: Adequacy MQM and Fluency MQM statistics across different evaluation sets. The reported statistics include the variance of system-level scores, the (S)PA of the scores with respect to All MQM, the F-statistic, and the B transformation of the difference between the p-values of Adequacy MQM and Fluency MQM. The last two are explained in §4.2.

differ primarily in adequacy, while performing similarly in fluency. This naturally leads to higher system-level variance in Adequacy MQM than in Fluency MQM.

Both of the above factors contribute to the variances of system-level Adequacy MQM and Fluency MQM, which in turn lead to the bias of the meta-evaluation. Therefore, we say that the meta-evaluation bias consists of two components: *intrinsic bias* and *extrinsic bias*.

The intrinsic bias reflects translation experts' beliefs about translation errors. Therefore, we retain intrinsic bias in this study. However, whether the extrinsic bias should be eliminated, retained, or partially kept is debatable and depends on the application. In our study, we aim to avoid it in order to gain a clearer understanding of the adequacy–fluency tradeoff at the evaluation level. In other contexts, one could argue that adequacy assessment should have greater influence on meta-evaluation outcomes, even at the cost of reduced fluency influence. In both cases, the extrinsic bias must first be measured separately from the intrinsic bias.

4.2 Measuring the Extrinsic Bias

To study the extrinsic bias, we propose to compare the F-statistics (Weir and Cockerham, 1984) for Adequacy MQM and Fluency MQM. In each case, the F-statistic is formulated as follows:

$$\text{F-statistic} = \frac{\text{between-system variation}}{\text{within-system variation}} \quad (3)$$

where the numerator and denominator are

$$\text{between-system variation} = \sum_{i=1}^{K} \frac{N \cdot (\bar{s}_i - \bar{s})^2}{K-1} \quad (4)$$

$$\text{within-system variation} = \sum_{i=1}^{K} \sum_{j=1}^{N} \frac{N \cdot (s_{i,j} - \bar{s}_i)^2}{N-K} \quad (5)$$

In our scenario, $s_{i,j}$ is the Adequacy MQM or Fluency MQM score of the candidate translation given by the ith system for the jth segment, $\bar{s}_i$ is the average of all the scores by the ith system, and $\bar{s}$ is the average of all the scores. N and K are the numbers of segments and systems, respectively.

The F-statistic is effective at distinguishing extrinsic variation from intrinsic variation because of how its numerator and denominator are defined. On one hand, the intrinsic variation can be regarded as a multiplicative constant within the s variables; it appears in both the numerator and denominator of Eqn. (3) and thus cancels out. On the other hand, the F-statistic captures extrinsic variation, which is due to the selected systems in the meta-evaluation. It does so by normalizing between-system variation with within-system variation.

Notice that the F-statistic depends on N and K, which makes it difficult to interpret. To this end, we adopt the ANOVA framework (Heiman, 2001) and use p-values to quantify extrinsic variation. We perform this analysis separately for Adequacy MQM and Fluency MQM.

We assume that, within each translation system, scores are independent and normally distributed around that system's mean, and that all systems share the same variance[5]; i.e., $s_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$. Under the null hypothesis that all systems have the same mean ($\mu_1 = \mu_2 = \cdots$), the F-statistic in Eqn. (3) follows an F distribution (Johnson et al., 1995) with degrees of freedom $K-1$ and $N-K$. The right-tail p-values are then given by

[5]ANOVA is robust to moderate variance inequality, so this assumption need only hold approximately (Lowry, 1998–2023). For completeness, we report in Appendix D the results using the variance-inequality-tolerant alternative ANOVA (Welch, 1951), which yield the same conclusions; We prefer the standard approach for its simplicity and greater numerical stability.

	Original	Synthesized by		Adequacy MQM			Fluency MQM			
		Adequacy	Fluency	F-stat	PA	SPA	F-stat	PA	SPA	$B(\overline{\Delta p})$
1	✓	–	–	38.5	0.96	0.94	7.7	0.76	0.76	0.120^{A}
2	–	✓	–	213.3	0.94	0.95	2.5	0.53	0.54	0.590^{A}
3	–	–	✓	7.5	0.62	0.71	165.1	0.82	0.78	0.450^{F}
4	✓	✓	–	111.9	0.95	0.95	4.9	0.61	0.62	0.130^{A}
5	✓	–	✓	21.7	0.80	0.83	75.5	0.73	0.73	0.030^{F}
6	–	✓	✓	93.6	0.84	0.87	72.4	0.59	0.60	$\mathbf{0.005^{A}}$
7	✓	✓	✓	72.7	0.88	0.88	48.8	0.63	0.64	$\mathbf{0.005^{A}}$

Table 3: Adequacy MQM and Fluency MQM statistics for different meta-evaluation setups, along with the B transformation of the difference between their corresponding p-values, macro-averaged across evaluation sets. Each meta-evaluation setup includes one or more system sets drawn from the original, synthesized-by-adequacy, and synthesized-by-fluency system sets.

$1 - \mathrm{CDF}(\text{F-statistic})$, where CDF denotes the cumulative distribution function of the F-distribution. Converting F-statistics to p-values puts results on a common probability scale, and makes the interpretation independent of N and K. We use this p-value as our measure of extrinsic variation.

As mentioned in §4.1, extrinsic variations of Adequacy MQM and Fluency MQM contribute to their respective variances. An asymmetry in these variations induces a bias of meta-evaluation toward adequacy or fluency, which we term *extrinsic bias*. We quantify this bias by

$$\Delta p = p_{\text{ Adequacy MQM}} - p_{\text{ Fluency MQM}} \quad (6)$$

where $p_{\text{ Adequacy MQM}}$ and $p_{\text{ Fluency MQM}}$ are the respective p-values. A positive Δp indicates bias toward adequacy, a negative Δp indicates bias toward fluency, and $\Delta p = 0$ indicates no bias between adequacy and fluency.

We observe that the range of Δp is typically very small (from 10^{-7} to 10^{-31} in Table 2), even in cases where severe bias is present (as shown in our later analysis). For presentation purposes, we report the degree of bias using $B(\Delta p) = \frac{1}{1-\log|\Delta p|}$, and append a special symbol to indicate which aspect dominates: A for adequacy ($\Delta p > 0$) and F for fluency ($\Delta p < 0$). Note that $B \in [0, 1]$, with a lower B indicating less bias.

We report the B values in Table 2, where we compare the influence of Adequacy MQM and Fluency MQM on the meta-evaluation in five evaluation datasets. Results show a consistent extrinsic bias toward adequacy. We argue that this behavior stems from the particular composition of translation systems in the WMT datasets, and may be controlled by carefully selecting or synthesizing systems with more variance in their Fluency MQM.

4.3 Reducing the Extrinsic Bias

In this subsection, we provide a method to reduce meta-evaluation extrinsic bias. This not only has practical values, but also allows us to better evaluate the adequacy–fluency bias of existing translation evaluation metrics (to be discussed in §5).

We accomplish this by considering additional pseudo-translation systems in the meta-evaluation. In particular, we synthesize two sets of translation systems: one exhibiting extreme variations in system-level Adequacy MQM, and the other in Fluency MQM. This design helps to control the variation in each dimension and thereby reduces extrinsic bias through controlled mixing of candidate translation systems.

Suppose we have K original translation systems. We synthesize K adequacy-oriented systems in the following way: for each segment, we assume that the kth synthesized system generates the kth most adequate translation (according to Adequacy MQM), among the original K translation systems' outputs.[6] It is easy to see that, given the available translation candidates, such K synthesized translate systems exhibit extreme variation in Adequacy MQM at the system level. Likewise, we synthesize K fluency-oriented systems, and in total, we have $3K$ translation systems for meta-evaluation. It is emphasized that our research does not require additional annotation effort, as the above synthesizing process utilizes readily available MQM scores.

Table 3 illustrates how synthesized systems help control the extrinsic bias. The table reports the macro-average of metrics across the five evaluation sets (§3.1).

[6]For segments with tied scores, we rank them randomly.

Rows 2 and 3 represent extreme configurations that maximize the F-statistic for Adequacy MQM and Fluency MQM, respectively. We see that they yield extreme B values, with Row 2 being biased toward adequacy and Row 3 toward fluency. This expected pattern confirms that (1) the B value effectively captures extrinsic bias arising from system selection, and (2) our synthesis method can control such a bias.

We observe that the setup in Row 5 has an overall bias toward adequacy (indicated by higher PA and SPA scores), although the extrinsic bias is toward fluency (indicated by the F annotation). This suggests the existence of intrinsic bias discussed in §4.1, further justifying the need of separately quantifying the intrinsic and extrinsic bias.

We also observe that Rows 6 and 7 yield the lowest B values, and we consider them the most balanced meta-evaluation setups, which will be used in our study of adequacy–fluency tradeoff at the evaluation level in §5.

4.4 The Effect of Meta-Evaluation Bias on Metric Comparisons

In the previous parts, we have demonstrated that the standard WMT meta-evaluation is biased toward adequacy, and we propose to carefully synthesize pseudo-translation systems to control this bias. In this part, we show how the meta-evaluation bias can influence the development of translation metrics.

Table 4 presents how the original WMT setup (Row 1 of Table 3) and *our balanced setup* (Row 6 of Table 3) rank various metrics, using PA and SPA as the meta-metrics.

It is interesting to examine "CometKiwi 22 XXL" and "MetricX (ours)" in detail, shown by the bold lines in Table 4. As seen, CometKiwi 22 XXL consistently and significantly outperforms MetricX (ours) under the original meta-evaluation, which is biased toward adequacy; however, this trend is reversed under our balanced meta-evaluation setup. In §5, we will show that CometKiwi 22 XXL has a considerable bias toward adequacy, whereas MetricX (ours) demonstrates a more balanced behavior. This comparison shows that a metric is favored if its bias in the adequacy–fluency spectrum aligns with that of the meta-evaluation; consequently, the development of translation metrics inherits the bias of the meta-evaluation. Our analysis highlights the importance of studying translation meta-evaluation imbalance.

Metric	Original setup PA	Original setup SPA	Our balanced setup PA	Our balanced setup SPA
AdequacyX	0.881 4	0.866 4	0.847 1	0.833 1
AdequacyX QE	0.885 3	0.873 1	0.834 2	0.824 5
FluencyX	0.697 13	0.720 12	0.676 12	0.675 13
MetricX (ours)	**0.862** 7	**0.845** 8	**0.833** 3	**0.829** 3
MetricX QE (ours)	0.878 6	0.861 6	0.809 6	0.823 6
MetricX-24	0.880 5	0.866 3	0.820 5	0.831 2
MetricX-24 QE	0.888 2	0.867 2	0.820 4	0.827 4
Comet 22	0.850 8	0.847 7	0.800 8	0.804 8
CometKiwi 22	0.813 9	0.802 9	0.782 9	0.772 9
CometKiwi 22 XXL	**0.889** 1	**0.862** 5	**0.805** 7	**0.815** 7
Gemma 3 (4B)	0.701 12	0.754 10	0.652 13	0.736 10
BLEU (sent. level)	0.726 11	0.712 13	0.726 10	0.689 12
ChrF (sent. level)	0.739 10	0.731 11	0.722 11	0.721 11

Table 4: Meta-evaluation of translation metrics using the original WMT setup and our balanced setup. Decimal numbers denote the PA and SPA scores, macro-averaged across our five evaluation sets; blue squares indicate the corresponding rankings. We bold MetricX (ours) and CometKiwi 22 XXL, as they are discussed in detail in the main text. Note that the results here are not identical to those in the WMT reports as we are reporting the average results across specific evaluation sets.

Evaluation Set	Concordance	Discordance
En–De'23	136 (49%)	140 (51%)
Zh–En'23	124 (29%)	311 (71%)
En–De'24	282 (50%)	279 (50%)
En–Es'24	139 (43%)	186 (57%)
Ja–Zh'24	156 (51%)	149 (49%)

Table 5: Concordance and discordance between adequacy and fluency in our balanced setup, reported as system pair counts and percentages.

5 Analysis: Tradeoff at the Evaluation Level

In this section, we analyze the adequacy–fluency tradeoff at the evaluation level and determine the bias of each metric within this tradeoff. Note that the notion of bias considered here is related to but distinct from that in §4. We say a metric is biased toward adequacy or fluency if it ranks translation systems in a way that is more aligned with the ranking derived solely from that aspect, as measured by the corresponding MQM scores.

To study the aforementioned bias, we would like to separately meta-evaluate the adequacy and fluency of each metric, using the original WMT meta-evaluation setup and our balanced setup described in Row 6 of Table 3.[7] Notice that the latter is a

[7]Although Rows 6 and 7 in Table 3 have similar B values, we prefer the former because it is fully synthesized and has a

Metric	Original setup			Our balanced setup		
	Agreement in discordant cases		PA in concordant cases	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM		with Adequacy MQM	with Fluency MQM	
All MQM	86%	14%	100%	74%	26%	100%
Adequacy MQM	100%		100%	100%		100%
Fluency MQM		100%	100%		100%	100%
AdequacyX	65%	35%	93%	70%	30%	91%
AdequacyX QE	70%	30%	93%	69%	31%	91%
FluencyX	22%	78%	84%	44%	56%	77%
MetricX (ours)	58%	42%	93%	69%	31%	90%
MetricX QE (ours)	61%	39%	95%	66%	34%	89%
MetricX-24	64%	36%	94%	70%	30%	89%
MetricX-24 QE	66%	34%	94%	70%	30%	89%
Comet 22	67%	33%	90%	73%	27%	86%
CometKiwi 22	72%	28%	84%	76%	24%	81%
CometKiwi 22 XXL	78%	22%	92%	73%	27%	87%
Gemma 3 (4B)	38%	62%	81%	47%	53%	76%
BLEU (sentence level)	71%	29%	74%	59%	41%	79%
ChrF (sentence level)	71%	29%	76%	65%	35%	77%

Table 6: PA breakdown for the original setup and our balanced setup, macro-averaged across five evaluation sets. For each metric, the agreements with Adequacy MQM and Fluency MQM sum to 1, by formulation and because ties between system-level Adequacy MQM and Fluency MQM scores are rare.

synthetic setup, which may not be representative of real-world situations. It is included only for our adequacy–fluency analysis.

From the data statistics in Table 5, we observe an even stronger disagreement between adequacy- and fluency-based system rankings under the balanced setup than under the original setup (Table 1). This is expected, as the balanced setup synthesizes the most- and least-adequate systems, which are unlikely to correspond to the most- and least-fluent systems, and vice versa.

Despite the significant discordance between Adequacy MQM and Fluency MQM, we also observe a large number of concordant cases (i.e., one system outperforms another in both Adequacy MQM and Fluency MQM). This raises a challenge when we separately analyze adequacy and fluency. To address this, we design three experimental protocols, as follows.

5.1 PA Breakdown

In this part, we design an experimental protocol using PA to separately analyze how a metric assesses adequacy and fluency, without being misled by the concordant cases. Since PA counts binary agreements, we may simply exclude concordant cases from our adequacy/fluency analysis.

Specifically, we first split our system pairs into two disjoint subsets based on whether their Adequacy MQM and Fluency MQM are concordant or discordant. Then, for each metric we report:

- PA in concordant cases, measuring the metric's general performance;
- PA with respect to Adequacy MQM in discordant cases, which measures the metric's bias toward adequacy (thus we call it "agreement with Adequacy MQM"); and
- PA with respect to Fluency MQM in discordant cases, which is called "agreement with Fluency MQM."

In Table 6, we present the results macro-averaged[8] across different evaluation sets.[9]

As seen, both BLEU and ChrF achieve more agreement with Adequacy MQM than Fluency MQM. This provides empirical evidence for the belief in the literature that BLEU and ChrF lean toward adequacy (Flamich et al., 2025). On the other hand, FluencyX and Gemma 3 exhibit a stronger bias toward fluency, which aligns with their design. Moreover, All MQM shows a strong alignment with Adequacy MQM in the original setup and a weaker alignment in our balanced setup, confirming our claims in §4 that current WMT meta-evaluation is biased toward adequacy and that such bias is partially mitigated in our balance meta-evaluation setup.

clearer construction. For completeness, we also report results for the setup in Row 7 in Appendix E.

[8] In our preliminary experiments, we also analyzed the micro-averaged results, which led to the same conclusions as those obtained here.

[9] We report results for individual evaluation sets under the original setup in Appendix F, where we observe diverse behaviors by the metrics. This variation may be due to the noise of metric performance, which tends to be smoothed out when averaged across evaluation sets.

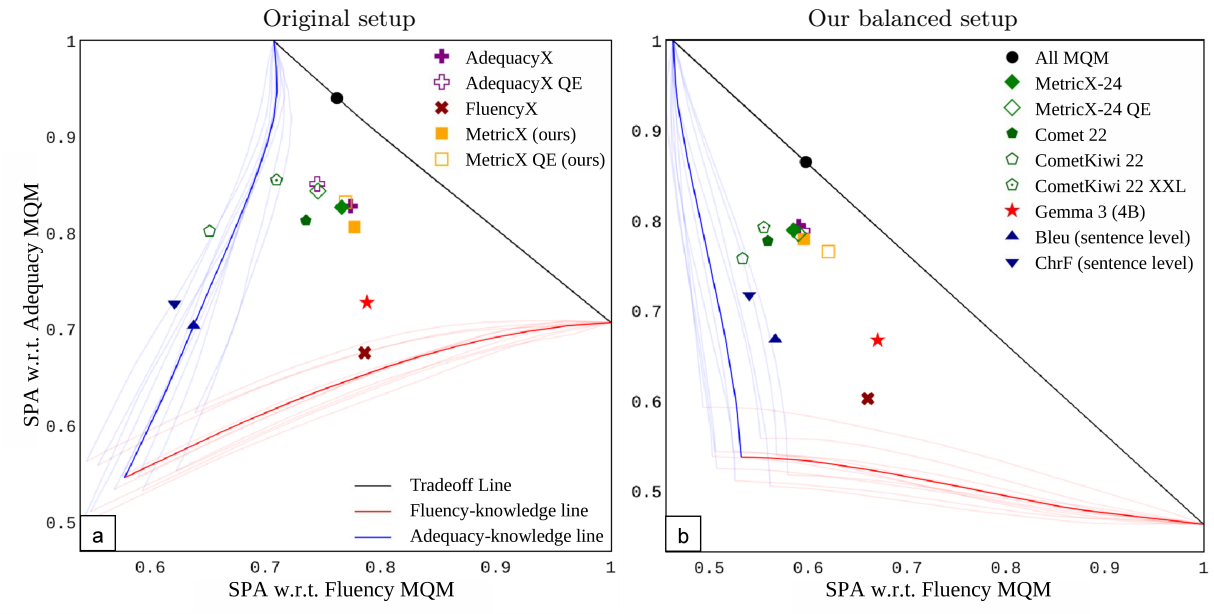


Figure 1: SPA plane using (a) the original setup and (b) our balanced setup. Legends are shared between the two figures. Each shadow line represents the combination of Adequacy MQM or Fluency MQM with a specific random noise instance. The solid red and blue lines are the average of the shadow lines.

Interestingly, FluencyX and Gemma 3 show some alignment with adequacy, despite lacking access to the source texts. This can be explained by two possibilities: (1) These metrics are imperfect and fail to align entirely with Fluency MQM, although designed to do so. (2) When the metrics are trained, they have learned strong prior for adequacy errors (e.g., due to the adequacy–fluency correlation of training segments). This aligns with the findings of Perrella et al. (2024), who show that metrics based solely on candidate translations or source texts can achieve high performance.

For most other metrics, we find that they exhibit stronger alignment with Adequacy MQM than with Fluency MQM. Among them, the MetricX variants display a relatively more balanced behavior than other metrics, including the Comet variants.

Finally, we would like to point out that our analysis sheds light on the position of different evaluation metrics within the adequacy–fluency tradeoff. However, whether a metric should achieve 50%–50% balance or mimicking All MQM (which is highly biased toward adequacy but often treated as the evaluation ground truth) is a task-specific design choice.

5.2 SPA Plane

We also aim to conduct an adequacy–fluency analysis using SPA. However, we cannot mirror the PA breakdown in §5.1, because SPA lacks a binary notion of concordance versus discordance. Instead, we perform qualitative visualization to demonstrate the adequacy–fluency tradeoff of a metric.

Specifically, we design a plot where the x-axis and y-axis represent SPA computed with respect to Fluency MQM and Adequacy MQM, respectively. Each metric is shown as a single point in this space. We then augment the plot with three sentinel lines:

- *Tradeoff line* (the black lines in Figure 1), representing the linear interpolation of Adequacy MQM and Fluency MQM. No system can surpass the tradeoff line.
- *Adequacy-knowledge line* (the blue lines in Figure 1), derived from linear interpolation of Adequacy MQM and random scores (i.e., uniform in the range of Adequacy MQM scores for each segment). The interpolation is accomplished by weighted average of the scores. Points on this line achieve fluency scores solely due to the correlation between Adequacy MQM and Fluency MQM.
- *Fluency-knowledge line* (the red lines in Figure 1), derived from linear interpolation of Fluency MQM and the random score.

The latter two lines are averaged from 10 computations with different random score instances.

These three lines form a triangle-like shape with vertices at Adequacy MQM, Fluency MQM, and pure random noise. A metric's closeness to the tradeoff line indicates its general quality, while its closeness to the other two lines reveals its bias toward adequacy or fluency.

Metric	Unnormalized sensitivity		Normalized sensitivity	
	Adequacy	Fluency	Adequacy	Fluency
All MQM	0.995	0.998	0.882	0.353
Adequacy MQM	1.000	0.000	1.000	0.000
Fluency MQM	0.000	1.000	0.000	1.000
AdequacyX	0.211	0.146	0.667	0.184
AdequacyX QE	0.143	0.084	0.474	0.111
FluencyX	0.010	0.018	0.202	0.145
MetricX (ours)	0.125	0.100	0.644	0.206
MetricX QE (ours)	0.081	0.070	0.521	0.180
MetricX-24	0.379	0.257	0.591	0.160
MetricX-24 QE	0.300	0.207	0.503	0.139
Comet 22	0.010	0.006	0.593	0.142
CometKiwi 22	0.007	0.003	0.415	0.071
CometKiwi 22 XXL	0.017	0.010	0.489	0.122
Gemma 3 (4B)	0.037	0.101	0.212	0.231
BLEU (sent. level)	1.201	0.755	0.383	0.099
ChrF (sent. level)	1.088	0.657	0.149	0.090

Table 7: Normalized and unnormalized sensitivity against adequacy and fluency.

Figure 1 presents the results, also macro-averaged across the evaluation sets in §5.1. The black line illustrates the severity of the adequacy–fluency tradeoff in translation evaluation (given a certain meta-evaluation setup): a larger top-right area indicates a more severe tradeoff, as this area is not reachable. As shown, our balanced setup exhibits a more severe tradeoff, which is consistent with Table 5. Moreover, our balanced setup has a larger angle between the blue and red lines, indicating a lower correlation between the two aspects.

Regarding specific metrics, we observe that FluencyX and Gemma 3 lie close to the fluency boundary (red line), while other metrics (namely, BLEU, ChrF, Comet, and MetricX) are biased toward adequacy. In particular, Comet variants are more biased than MetricX variants. The SPA results are consistent with the analysis using PA (§5.1).

5.3 Sensitivity Analysis

In §5.1 and §5.2, we illustrate the position of each metric in the adequacy–fluency tradeoff by their system-level ranking prediction. Here, we aim to analyze how each metric reacts to an adequacy or fluency error.

We do this by controlling one aspect while varying the other. Take adeqacy as an example. Given a source segment, we consider pairs of translation candidates that share the same Fluency MQM but differ in Adequacy MQM. For a metric, we report $\frac{\Delta \text{ metric score}}{\Delta \text{ Adequacy MQM}}$, averaged over all our translation pairs. This measures the expected change in the metric score per one-point difference in Adequacy MQM.

We also report a normalized version of the above measure by multiplying it with $\frac{\sum \sigma(\text{Adequacy MQM})}{\sum \sigma(\text{metric score})}$, where $\sigma(\cdot)$ is the standard deviation over different candidate translations given a segment, and the sum is taken over all segments. This scaling enables meaningful comparison across metrics with different output scales.

Table 7 reports the results for both adequacy and fluency. Here, we can interpret the bias of a metric by comparing the sensitivity to adequacy and that to fluency. Generally, our findings in previous subsections are also observed in this experiment, except for the normalized sensitivity of FluencyX, which requires further investigation.

Overall, this section provides three analysis protocols (PA Breakdown, SPA Plain, and Sensitivity Analysis) to study the position of different evaluation metrics within the adequacy–fluency tradeoff. Our key findings in this section include: (1) Most of the translation metrics are biased toward adequacy, and (2) For the commonly used MetricX and Comet metrics, the former exhibits a more balanced behavior than the latter. Consistent observations in different protocols cross-validate the reliability of our findings.

6 Conclusion

In this work, we investigate the adequacy–fluency tradeoff in translation. While this tension is well-documented at the level of translation output (Flamich et al., 2025), we show that it also manifests severely at the levels of evaluation and meta-evaluation.

Through empirical analysis of WMT meta-evaluation protocols, we uncover a systematic bias toward adequacy, driven by the composition of meta-evaluation datasets. We also propose a method that can control the balance between adequacy and fluency.

We further analyze the placement of popular translation evaluation metrics along the adequacy–fluency tradeoff and find that most metrics lean toward adequacy.

We argue that the adequacy–fluency tradeoff is a critical yet under-recognized matter in translation evaluation and meta-evaluation. While we do not take a stance on how to deal with such bias, our primary contribution is to raise awareness of this matter within the community.

We discuss future work in Appendix G.

Limitations

Our work provides an extensive investigation of adequacy–fluency trade-offs in the evaluation and meta-evaluation of machine translation. However, it also has limitations.

First, our work is based on MQM data, which includes human evaluation that is inherently noisy. It places limits on the generality of the conclusion we draw in this paper. To mitigate this, we report results macro-averaged over five translation datasets. The results on individual dataset are considerably noisier, as shown in Appendix F.

Second, this study covers only a limited set of language pairs and systems, owing to the limited availability of MQM data. Expanding the evaluation data would yield more robust conclusion.

Third, we only analyze extrinsic bias (caused by the composition of translation systems in meta-evaluation), while not addressing the intrinsic bias (caused by the design of the MQM framework). Studying the intrinsic bias is highly impactful and warrants further efforts.

References

Sweta Agrawal, George Foster, Markus Freitag, and Colin Cherry. 2021. Assessing reference-free peer evaluation for machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1171.

Bogdan Babych and Anthony Hartley. 2008. Sensitivity of automated MT evaluation metrics on higher quality MT output: BLEU vs task-based evaluation methods. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 2133–2136.

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating MT in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Shweta Chauhan and Philemon Daniel. 2023. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, page 12663–12717.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929.

Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. You cannot feed two birds with one score: The accuracy–naturalness tradeoff in translation. *arXiv preprint arXiv:2503.24013*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 Metrics Shared Task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 Metrics Shared Task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Gary W Heiman. 2001. *Understanding Research Methods and Statistics: An Integrated Introduction for Psychology*. Houghton, Mifflin and Company.

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey. *arXiv preprint arXiv:2501.12011*.

Norman L Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. 1995. *Continuous Univariate Distributions*, volume 2. John Wiley & Sons.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494.

Alon Lavie and Michael J Denkowski. 2009. The METEOR metric for automatic evaluation of machine translation. *Machine Translation*, 23(2):105–115.

Richard Lowry. 1998–2023. *Concepts & Applications of Inferential Statistics*. Online Book.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers)*, pages 62–90.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 Metrics Shared Task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51.

Valerie Mariana, Troy Cox, and Alan Melby. 2015. The multidimensional quality metrics (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation*, pages 137–161.

Marianna Martindale and Marine Carpuat. 2018. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 Metrics Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 16216–16244.

John R. Pierce and John B. Carroll. 1966. *Language and Machines: Computers in Translation and Linguistics*. National Academy of Sciences/National Research Council, USA.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: An unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 722–729.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234.

B. S. Weir and C. Clark Cockerham. 1984. Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

B. L. Welch. 1951. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336.

William J. Welch. 1990. Construction of permutation tests. *Journal of the American Statistical Association*, 85(411):693–698.

John S. White and Theresa A. O'Connell. 1993. Evaluation of machine translation. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training neural machine translation with semantic similarity. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

A Adequacy MQM and Fluency MQM Categorization

As mentioned in §3.1, we consider Adequacy MQM and Fluency MQM separately, following the categorizations provided in Flamich et al. (2025). Table 8 shows the categorization used for En–De, Ja–Zh, and Zh–En. Specially, En–Es follows the categorization in Table 9.

B Dataset Statistics

Table 10 summarizes key statistics for each evaluation dataset in our study.

C PA vs. SPA

In the main paper, we mention that PA and SPA could behave differently, and that it is necessary to consider both PA and SPA in our study. In this appendix, we use a toy example to illustrate the potential disagreement of PA and SPA when assessing whether a metric is biased toward Adequacy MQM or Fluency MQM.

Table 11 shows the details of our example. In this case, PA suggests that the metric is biased toward adequacy, because PA only judges based on the sign. However, SPA suggests the opposite, as it concerns the closeness of the score difference. See §2.2 for details.

Therefore, we use both of the metrics, and design dedicated experiments for them. Luckily, our findings are generally consistent under PA and SPA, suggesting that they are reliable meta-metrics in our study.

D Measuring Extrinsic Bias without Equal-Variance Assumption

In §4.2, for simplicity we assume that all translation systems share the same variance in their Adequacy MQM scores and the same variance in their Fluency MQM scores. In this part, we drop this assumption and use Welch (1951)'s ANOVA approach (which does not require the equal-variance assumption) to calculate F-statistics and B values.

Table 12 reports our results. The findings are consistent with those reported in Table 2 under the equal-variance assumption. We nevertheless prefer the standard approach presented in the main text due to its greater simplicity and numerical stability. Note that we observe undefined p-values in our preliminary experiments for some synthetic datasets in §4.3, if we do not have the assumption.

E Analysis Results for Setup 7

In this section, we present the results of PA Breakdown (§5.1) and SPA Plane (§5.2) based on Setup 7 in Table 3. We include this setup for the sake of completeness, as it closely matches Row 6 (used in our main analysis) in terms of B values. Table 13, Table 14, and Figure 2 present the concordance/discordance ratio, PA Breakdown, and SPA Plane results, respectively, for Setup 7. All our findings based on Setup 6 also hold in this setup.

F Adequacy–Fluency Tradeoff Results on Individual Evaluation Sets

In §5, we report results macro-averaged across five evaluation sets (of different language pairs). In this part, we present the results on each evaluation set.

In Table 15 and Figure 3, we report PA-Breakdown and SPA Planes results based on the original WMT setup, presented separately for each evaluation set. The results exhibit diverse behaviors across evaluation sets. Thus, we perform macro-average in our main paper. The noisy phenomenon requires further investigation.

G Future Work

Our study opens several avenues for future research, discussed below.

Accuracy errors	Fluency errors		Other
Accuracy/Addition	Fluency/Grammar	Style/Archaic or obscure word choice	Other
Accuracy/Creative Reinterpretation	Fluency/Inconsistency	Style/Bad sentence structure	Source issue
Accuracy/Gender Mismatch	Fluency/Punctuation	Style/Unnatural or awkward	
Accuracy/Mistranslation	Fluency/Register	Locale convention/Address format	
Accuracy/Omission	Fluency/Spelling	Locale convention/Currency format	
Accuracy/Source language fragment	Fluency/Text-Breaking	Locale convention/Time format	
Non-translation!	Terminology/Inconsistent	Terminology/Inappropriate for context	

Table 8: Flamich et al. (2025)'s categorization of MQM errors, used for En–De, Ja–Zh, and Zh–En translation directions.

Accuracy errors	Fluency errors		Other
Addition	Capitalization	Date-time format	Other
Agreement	Inconsistency	Lacks creativity	Source issue
Do not translate	Grammar	Measurement format	
Mistranslation	Number format	Punctuation	
MT hallucination	Register	Spelling	
Omission	Unnatural flow	Whitespace	
Untranslated	Word order	Wrong language variety	
Wrong named entity			
Wrong term			

Table 9: Flamich et al. (2025)'s categorization of MQM errors, used for the En–Es translation direction.

Year	2023	2023	2024	2024	2024
Language pair	En–De	Zh–En	En–De	En–Es	Ja–Zh
Mean All MQM	6.17	2.51	2.50	0.58	2.92
Mean Adequacy MQM	3.97	1.62	1.38	0.45	2.55
Mean Fluency MQM	1.99	0.89	1.10	0.13	0.35
Mean non-zero All MQM	10.02	5.20	4.95	3.00	6.00
Mean non-zero Adequacy MQM	9.39	6.21	5.63	3.91	6.38
Mean non-zero Fluency MQM	4.23	2.36	2.82	1.37	2.07
# segments	5520	1954	8766	8722	7840
# segments w/o errors (All MQM = 0)	1350	350	3901	6672	3842
# segments w/ errors (All MQM > 0)	4170	1604	4865	2050	3998
# segments w/ adequacy errors (Adequacy MQM > 0)	2919	891	2738	1393	3354
# segments w/ fluency errors (Fluency MQM > 0)	3149	1250	3462	849	1325

Table 10: Dataset statistics by language pair.

First, we reveal the potential imbalance in the translation meta-evaluation and highlight the importance of understanding this phenomenon (§4). We propose a statistical measure to quantify this imbalance and design a method to control it through data synthesis. However, there is room for improvement in both measuring the imbalance and debiasing the meta-evaluation. In particular, we are interested in exploring theoretical approaches to debiasing the meta-evaluation outputs by normalizing scores in a post hoc manner, without altering the datasets.

Second, we develop dedicated translation metrics, AdequacyX and FluencyX, to independently assess adequacy and fluency (§3.3). We aim to further improve this separation, as it facilitates adequacy–fluency analyses, especially in scenarios where MQM annotations are not available.

Third, as we are now in the era of large language models (LLMs), it is interesting to study the LLM-as-a-judge for translation through the lens of the adequacy–fluency tradeoff, for example, understanding and steering the bias of LLM-as-a-judge.

	Sys 1	Sys 2	Δ
Adequacy MQM	8.0	6.0	+2.0
Fluency MQM	6.0	6.5	−0.5
Metric	6.3	6.0	+0.3

closer

same sign

Table 11: A toy example illustrating potential disagreement between PA and SPA: PA prefers keeping the same sign, whereas SPA prefers the closer one. Here, all numbers are hypothetical for illustration purposes.

	F-statistic		$B(\Delta p)$
	Adequacy	Fluency	
En–De'23	28.9	8.4	0.07^A
Zh–En'23	84.1	11.9	0.04^A
En–De'24	13.0	9.0	0.04^A
En–Es'24	26.5	4.2	0.14^A
Ja–Zh'24	28.8	6.9	0.08^A
Macro-Avg	36.3	8.1	

Table 12: Welch (1951)'s F-statistics for Adequacy MQM and Fluency MQM, and the respective B-values, in the original meta-evaluation setup.

Evaluation Set	Concordance	Discordance
En–De'23	356 (57%)	274 (43%)
Zh–En'23	370 (37%)	620 (63%)
En–De'24	692 (54%)	583 (46%)
En–Es'24	375 (51%)	366 (49%)
Ja–Zh'24	399 (54%)	342 (46%)

Table 13: Concordance and discordance between Adequacy MQM and Fluency MQM in Setup 7, reported as system pair counts and percentages.

	Setup 7	
Metric	Agreement in discordant cases with Adequacy MQM / with Fluency MQM	PA in concordant cases
All MQM	76% / 24%	100%
Adequacy MQM	100%	100%
Fluency MQM	100%	100%
AdequacyX	67% / 33%	92%
AdequacyX QE	67% / 33%	92%
FluencyX	41% / 59%	79%
MetricX (ours)	66% / 34%	92%
MetricX QE (ours)	64% / 36%	91%
MetricX-24	67% / 33%	91%
MetricX-24 QE	67% / 33%	91%
Comet 22	69% / 31%	88%
CometKiwi 22	72% / 28%	83%
CometKiwi 22 XXL	71% / 29%	90%
Gemma 3 (4B)	46% / 54%	78%
BLEU (sentence level)	61% / 39%	78%
ChrF (sentence level)	64% / 36%	78%

Table 14: PA breakdown, macro-averaged across five evaluation sets, for Setup 7.

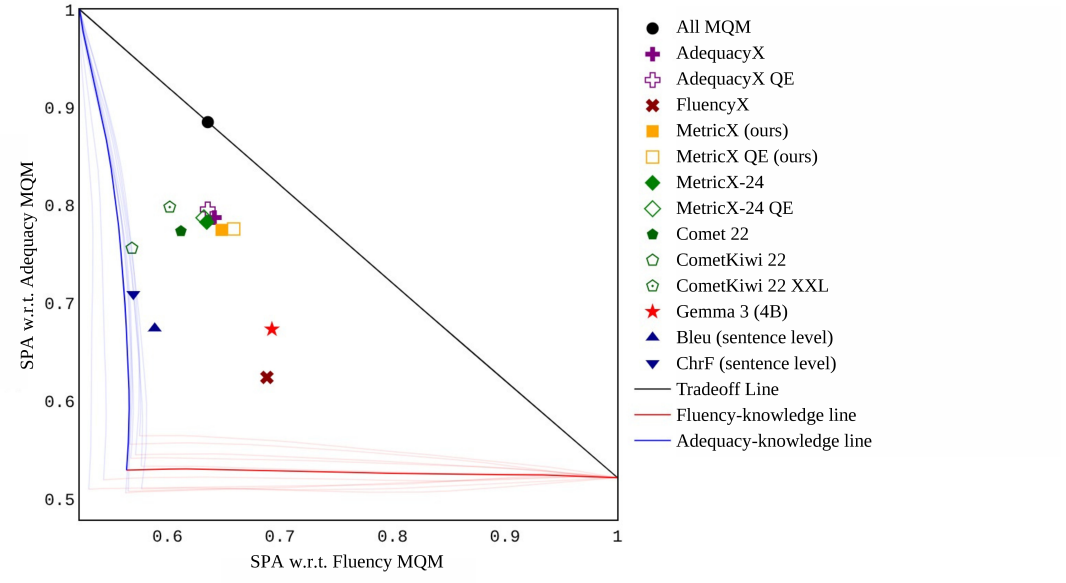


Figure 2: SPA plane using Setup 7, macro-averaged across five evaluation sets. Each shadow line represents the combination of Adequacy MQM or Fluency MQM with a specific random noise instance. The solid red and blue lines are the average of the shadow lines.

Metric	En-De 2023			Zh-En 2023		
	Agreement in discordant cases		PA in concordant cases	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM		with Adequacy MQM	with Fluency MQM	
All MQM	94%		100%	91%		100%
Adequacy MQM	100%		100%	100%		100%
Fluency MQM		100%	100%		100%	100%
AdequacyX	76%	24%	100%	77%	23%	94%
AdequacyX QE	88%		100%	71%	29%	96%
FluencyX	35%	65%	92%		91%	89%
MetricX (ours)	65%	35%	100%	66%	34%	87%
MetricX QE (ours)	71%	29%	100%	66%	34%	96%
MetricX-24	94%		98%	74%	26%	96%
MetricX-24 QE	94%		98%	71%	29%	96%
Comet 22	88%		98%	80%	20%	86%
CometKiwi 22	94%		96%	71%	29%	93%
CometKiwi 22 XXL	94%		98%	74%	26%	94%
Gemma 3 (4B)	65%	35%	90%		91%	73%
BLEU (sentence level)	88%		88%	80%	20%	77%
ChrF (sentence level)	88%		82%	71%	29%	80%

Metric	En-De 2024			En-Es 2024		
	Agreement in discordant cases		PA in concordant cases	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM		with Adequacy MQM	with Fluency MQM	
All MQM	55%	45%	100%	87%	13%	100%
Adequacy MQM	100%		100%	100%		100%
Fluency MQM		100%	100%		100%	100%
AdequacyX	42%	58%	92%	57%	43%	87%
AdequacyX QE	53%	47%	90%	70%	30%	87%
FluencyX	16%	84%	94%	26%	74%	71%
MetricX (ours)	45%	55%	94%	57%	43%	89%
MetricX QE (ours)	50%	50%	95%	65%	35%	89%
MetricX-24	42%	58%	92%	48%	52%	85%
MetricX-24 QE	53%	47%	97%	52%	48%	85%
Comet 22	58%	42%	95%	48%	52%	84%
CometKiwi 22	82%	18%	78%	57%	43%	69%
CometKiwi 22 XXL	74%	26%	89%	65%	35%	87%
Gemma 3 (4B)	47%	53%	82%	48%	52%	73%
BLEU (sentence level)	63%	37%	80%	43%	57%	53%
ChrF (sentence level)	71%	29%	80%	43%	57%	64%

Metric	Ja-Zh 2024		
	Agreement in discordant cases		PA in concordant cases
	with Adequacy MQM	with Fluency MQM	
All MQM	100%		100%
Adequacy MQM	100%		100%
Fluency MQM		100%	100%
AdequacyX	75%	25%	93%
AdequacyX QE	70%	30%	91%
FluencyX	25%	75%	72%
MetricX (ours)	60%	40%	97%
MetricX QE (ours)	55%	45%	95%
MetricX-24	60%	40%	97%
MetricX-24 QE	60%	40%	95%
Comet 22	60%	40%	88%
CometKiwi 22	55%	45%	86%
CometKiwi 22 XXL	85%	15%	90%
Gemma 3 (4B)	20%	80%	86%
BLEU (sentence level)	80%	20%	72%
ChrF (sentence level)	80%	20%	76%

Table 15: PA breakdown, per evaluation set, based on the original WMT setup.

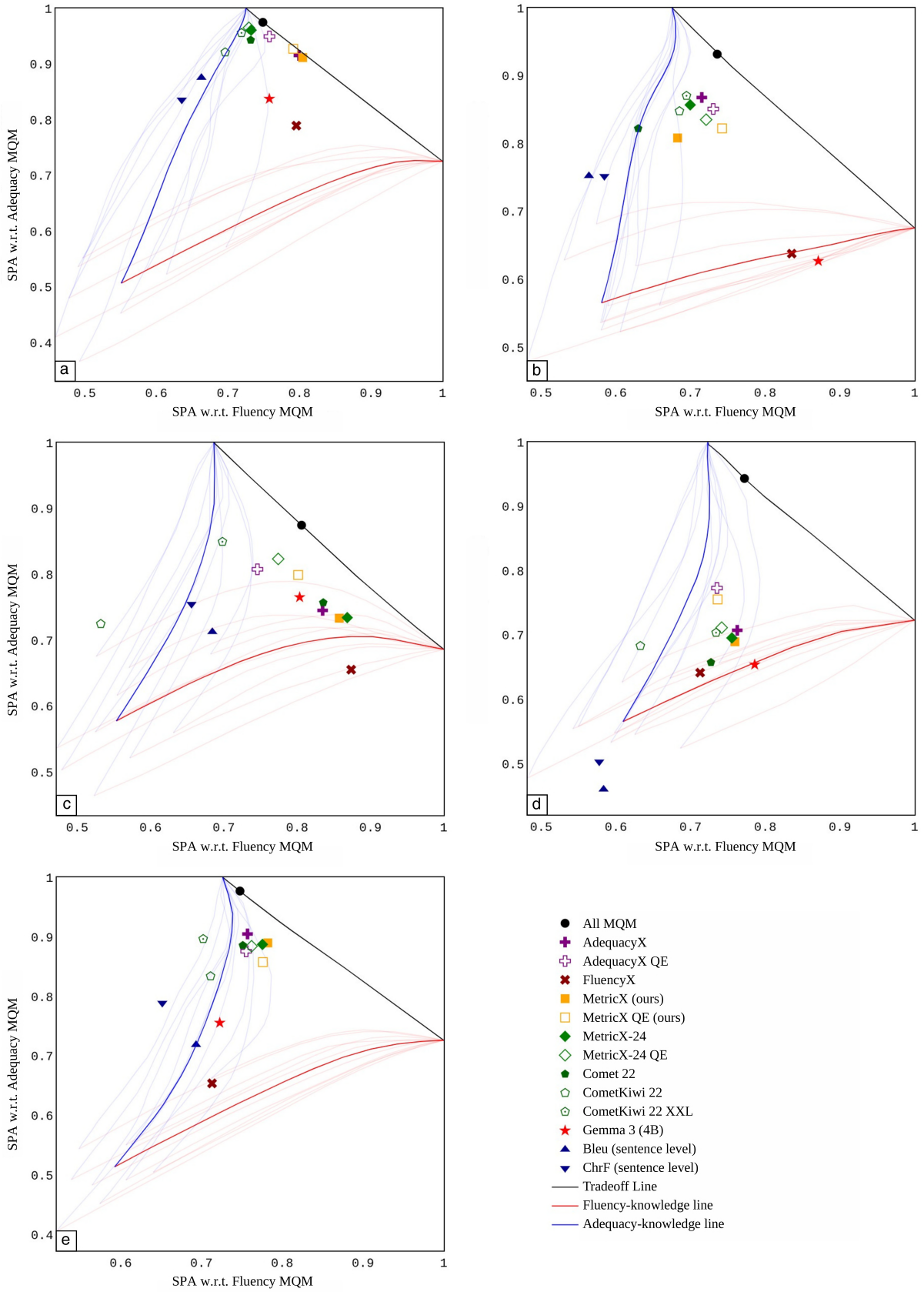


Figure 3: SPA plane based on the original WMT setup for (a) En–De 2023, (b) Zh–En 2023, (c) En–De 2024, (d) En–Es 2024, and (e) Ja–Zh 2024. The legend is shared for all the figures.

DocHPLT: A Massively Multilingual Document-Level Translation Dataset

Dayyán O'Brien[★,♚] **Bhavitvya Malik**[★,♚] **Ona de Gibert**[♞]
Pinzhen Chen[♚] **Barry Haddow**[♚] **Jörg Tiedemann**[♞]
[♚]University of Edinburgh [♞]University of Helsinki
{dayyan.obrien,bmalik2,pinzhen.chen,bhaddow}@ed.ac.uk
{ona.degibert,jorg.tiedemann}@helsinki.fi

Abstract

Existing document-level machine translation resources are only available for a handful of languages, mostly high-resourced ones. To facilitate the training and evaluation of document-level translation and, more broadly, long-context modeling for global communities, we create DocHPLT, the largest publicly available document-level translation dataset to date. It contains 124 million aligned document pairs across 50 languages paired with English, comprising 4.26 billion sentences. By adding pivoted alignments, practitioners can obtain 2500 additional pairs not involving English. Unlike previous reconstruction-based approaches that piece together documents from sentence-level data, we modify an existing web extraction pipeline to preserve complete document integrity from the source, retaining all content, including unaligned portions. After our preliminary experiments identify the optimal training context strategy for document-level translation, we demonstrate that LLMs fine-tuned on DocHPLT substantially outperform off-the-shelf instruction-tuned baselines, with particularly dramatic improvements for under-resourced languages. We open-source the dataset under a permissive license, providing essential infrastructure for advancing multilingual document-level translation.

1 Introduction

The field of natural language processing (NLP) is shifting its focus toward end-to-end, complex tasks, including the domain of machine translation. This increases the demand for techniques and resources beyond the sentence level, with document-level machine translation (DocMT) being a prime example (Maruf and Haffari, 2018; Zhang et al., 2018; Agrawal et al., 2018; Huo et al., 2020). While there is not a single definition of a document, DocMT requires models to translate more than one sentence as a coherent unit rather than isolated segments. This approach is necessary for handling various discourse phenomena: *anaphora, deixis, ellipsis, discourse connectives, grammatical and lexical cohesion* (Maruf et al., 2021), which sentence-level translation typically loses (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2018). Recent long-context large language models (LLMs) are well-suited for this task, as they are usually pre-trained to process thousands of tokens at a time. However, DocMT remains largely unexplored or untested for most languages due to a simple but significant problem: we lack document-level parallel data for both model building and evaluation.

Historically, parallel corpora were mostly constructed in a sentence-oriented manner, using a pipeline that split the text into sentences, aligned them, and then discarded unaligned and multiply-aligned sentences. While a handful of language pairs have some document-level MT resources, the majority of languages have none. This creates two related problems at once: we cannot build DocMT for these languages, and we cannot evaluate DocMT properly. As NLP research moves toward more end-to-end, context-aware applications, this data gap means that most languages get left behind.

We tackle this problem by extracting parallel documents from large web crawls, but our methodology differs from the majority of previous efforts that reconstruct data from sentence pairs after the fact. Instead, we modify the web extraction pipeline itself to preserve document structure from the beginning, retaining documents in their entirety with all original context and non-parallel text. For each language pair, we deliver the aligned documents along with quality-scored sentence alignments and alignment density metrics.

Our effort yielded DocHPLT, a large multilingual document-level translation dataset covering 50 language pairs with English, listed in Appendix A.

★Equal contribution. Public access to DocHPLT: https://huggingface.co/datasets/HPLT/DocHPLT.

Proceedings of the Tenth Conference on Machine Translation, pages 286–300
November 8-9, 2025

The resulting corpus contains 87.8 million documents in English and 50 other languages, 124 million aligned document pairs, and 4.26 billion sentences. A highlight of our work is the focus on medium- and low-resource languages that previous DocMT datasets have overlooked. Practitioners can also use English as a pivot to align up to 2500 extra non-English pairs, expanding the dataset's usefulness beyond English-centric translation.

Using DocHPLT, we conduct extensive experiments with different modeling methods for LLM-based document-level translation. We first try different context sizes for LLM fine-tuning: 1) full document-to-document training with loss calculated on the entire target; and 2) chunk-based training with loss computed on individual segments. These experiments determine the optimal context granularity for our subsequent work. Then, in addition to prompting off-the-shelf instruction-tuned large language models (LLMs) as a baseline, we run monolingual and multilingual fine-tuning using DocHPLT, tested on both seen and unseen languages. The usefulness of our data is reflected empirically: LLMs fine-tuned on our data consistently outperform prompting baselines, showing that practitioners can gain strong performance in DocMT for languages often considered "unsupported" in the machine translation research community.

In summary, our contributions, centred around the DocHPLT resource, are as follows:

- **Scale and diversity:** DocHPLT is the **largest** publicly available document-level translation resource: 124M document pairs for 50 languages paired with English, totaling 4.26B sentences, with extensive medium- and low-resource coverage.
- **Document-first approach**: Instead of piecing together documents from aligned sentence pairs, we preserve complete documents with original structure and unaligned text, enriched with quality metrics such as alignment density and sentence pair-level scores.
- **Empirical validation**: Through LLM experiments on both the internal test set and WMT24++, we establish baselines, test different training strategies, and demonstrate gains in DocMT using our data.

2 Related Work

2.1 Document-Level Translation

Document-level translation aims to process an entire document as a coherent unit, rather than processing each sentence independently. This paradigm leverages the ability of modern neural architectures, lately LLMs, to handle long context, making it particularly effective for capturing document-level discourse structures. Recent work has demonstrated that going beyond sentence-level translation is essential for handling discourse phenomena such as coreference resolution (Müller et al., 2018; Bawden et al., 2018; Voita et al., 2018). The development of dedicated document-level benchmarks further reflects this growing interest in evaluating MT systems in context (Guillou and Hardmeier, 2016; Jwalapuram et al., 2020; Wicks and Post, 2023; Fernandes et al., 2023).

Moreover, a variety of modeling strategies have been proposed for DocMT (Tiedemann and Scherrer, 2017; Maruf and Haffari, 2018; Zhang et al., 2018; Agrawal et al., 2018; Sun et al., 2022), and more recent works adapt LLM-based architectures (Wang et al., 2023; Petrick et al., 2023; Wu et al., 2024; Jin et al., 2024; Ramos et al., 2025; Hu et al., 2025). However, there is still no standard practice in training to ensure effective context handling or in assessing document-level translation. Also, performance gains over strong sentence-level baselines remain inconsistent and not clearly attributable to effective context utilization (Kim et al., 2019). In this work, we try out various context sizes in LLM fine-tuning to establish effective training strategies on DocHPLT.

2.2 Document-Level Translation Data

Although there have been several massive-scale parallel corpus mining efforts (Bañón et al., 2020; El-Kishky et al., 2020; Schwenk et al., 2021; de Gibert et al., 2024; Burchell et al., 2025), document-level data remain limited in size and scope, particularly when extending beyond English-centric or high-resource languages. Moreover, there is little agreement on what constitutes a "document"; definitions vary widely across studies, ranging from short paragraphs to entire articles or books. This lack of standardization, combined with the scarcity of large-scale multilingual document-level corpora, motivates the need for more diverse resources such as the one we present in this work. Before illustrating our data methodology, we survey two typical

methods used in creating document-level translation data.

Reconstruction-based A common strategy to obtain document-level parallel data is to reconstruct from existing sentence-level data. Notably, the sentence-level ParaCrawl (Bañón et al., 2020) has been widely used as a seed for this purpose. Al Ghussin et al. (2023) extracted English–German parallel paragraphs from ParaCrawl, although these are not full-document units. ParaDocs (Wicks et al., 2024) recovered document-level data from ParaCrawl, News Commentary (Kocmi et al., 2023), and Europarl (Koehn, 2005) for German, French, Spanish, Italian, Polish, and Portuguese, all paired with English. Similarly, Pal et al. (2024) released a large-scale reconstructed corpus for German, French, Czech, Polish, and Russian—again all paired with English, along with an open-source pipeline for extension to other languages.

Collection-based An alternative approach is to collect or create document-level parallel corpora directly from targeted sources from scratch. Earlier efforts include Europarl based on the proceedings of the European Parliament (Koehn, 2005) and OpenSubtitles from movie and TV subtitles (Lison and Tiedemann, 2016), but these essentially consist of "spoken" documents, where the former is divided into speeches and the latter into films/shows. Literary works have also been a popular origin. Jiang et al. (2022) introduced a Chinese–English corpus based on web novels, where each chapter, with a median of 30 sentences, is treated as a document. Thai et al. (2022) created PAR3 by aligning machine and human translations of 118 novels across 19 languages at the paragraph level. Jin et al. (2024) constructed JAM from 160 English-Chinese novel pairs with chapter-level alignment. More recently, Alabi et al. (2025) created AFRIDOC-MT, a document-level translation corpus sourced from IT news and health articles and manually translated from English. By covering Amharic, Hausa, Swahili, Yorùbá, and Zulu, it extends DocMT data to medium and lower-resourced languages. Wastl et al. (2025) scraped a Swiss online news outlet to create 20min-XD, a French-German dataset. Such data, directly derived from resources intended to be document-aligned, is high-quality but often limited by languages due to the coverage of the upstream source and/or the cost and effort required.

Key methodological differences in this work Our work gathers document translations from large web crawls, but differs fundamentally from reconstruction-based approaches. As explained later in Section 3, rather than piecing together documents from sentence pairs post-hoc, we modify the document alignment stage of the extraction pipeline to preserve complete document structure from the beginning. This document-first methodology ensures we retain all original content, including unaligned portions, positioning our approach as collection-based at the document level while leveraging existing text crawling and processing infrastructure.

3 DocHPLT

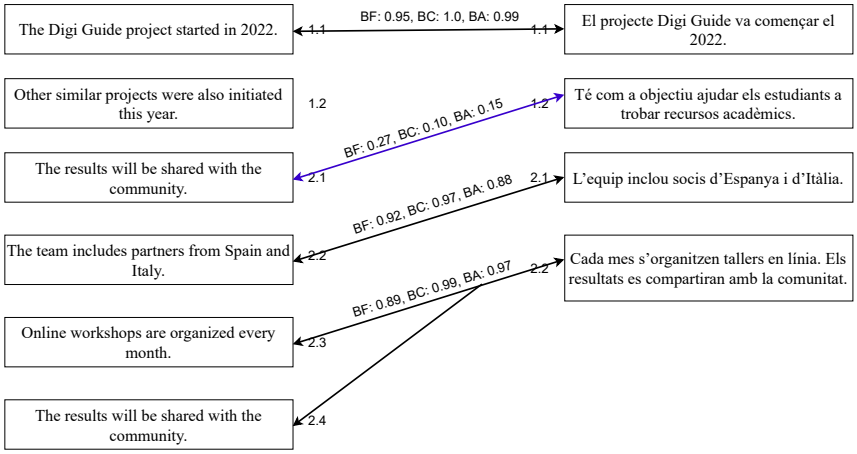
In this section, we explain how we modify and then apply an existing parallel sentence extraction pipeline from ParaCrawl to extract a document-level corpus from a large multilingual web crawl, HPLT.

3.1 Dataset Creation

The starting point for our dataset creation is 15TB of cleaned web documents derived from the Internet Archive[1] and CommonCrawl[2] released as version 2 of the HPLT corpus (Burchell et al., 2025). In the preparation of HPLT, the document text was extracted from HTML using Trafilatura (Barbaresi, 2021) and language-classified using openLID (Burchell et al., 2023). In this work, a *document* is defined as *the full text content retrieved from archive snapshots of a specific URL.*

To extract a parallel corpus of documents, we use a modified version of the ParaCrawl extraction pipeline (Bañón et al., 2020). The original pipeline is sentence-oriented, i.e. it produces a sentence-aligned corpus and discards unaligned sentences. But because the pipeline runs document alignment followed by sentence alignment in separate stages, we are able to intervene to produce document-oriented data. We extract and record each pair of aligned documents, then map the unfiltered sentence alignments back into their source documents. This document-first methodology ensures we retain all document content, even unaligned portions, to provide richer context than traditional parallel corpora.

[1] https://archive.org/
[2] https://commoncrawl.org/



BF: Bifixer, **BC**: Bicleaner, **BA**: BLEUalign. Higher values indicate better alignment and translation quality.

Figure 1: An example of good, bad (in blue), and multi-way alignments for English-Catalan docs.

Document structuring We transform each document into a hierarchical XML representation that preserves its internal structure. Paragraphs are split by newline characters and maintain their original boundaries, while the text of each paragraph is segmented into sentences using the Loomchild Segmenter (Miłkowski and Lipski, 2011). Every structural element receives a unique identifier with paragraphs, such as `<P id="4">`, and sentences, such as `<s id="4.3">`. This structured representation allows us to track alignments at both document and sentence levels while retaining all content from the original HPLT documents.

Content-based deduplication Since our initial collection contains multiple temporal snapshots of the same URLs, we implement a content-based deduplication strategy. First, within each language-specific collection, we remove duplicates, using the URL together with the full text as the key. This ensures we keep only unique document versions for each URL. Second, we perform global deduplication, based on the URL and text as the key, across all English documents from the 50 language pairs, consolidating them into a single collection. This is necessary because the same English document may appear in multiple language pairs (e.g., the same English page aligned to both Basque and Catalan translations). After deduplication, we have a clean collection of unique documents for each of the 50 source languages and a single, unified collection for all English documents, ensuring each unique document version appears exactly once while preserving all alignment relationships.

We deliberately preserve duplicate content across different URLs and retain near-duplicates within the same URL. This design choice maximizes research flexibility by allowing downstream users to apply filtering strategies suited to their particular use cases. Additionally, duplicate content from different URLs preserves valuable metadata, particularly the source URL, which may indicate different domains, publication contexts, or content distribution patterns. Near-duplicates also represent meaningful content variations such as updates, revisions, or editorial differences. Deduplicating content only for each language separately results in a 3.3% drop in our document count from the original DocHPLT corpus (see Appendix A).

Alignment verification and generation The ParaCrawl pipeline originally used MinHash (Broder, 1997) to deduplicate similar sentences, grouping them and assigning the same quality scores regardless of their source documents. We modify this step to remove MinHash deduplication entirely, instead maintaining all original texts and tracking which documents they came from. This allows us to preserve the complete document structure while still computing alignment quality scores—BLEUalign (Sennrich and Volk, 2010), Bicleaner, and Bifixer (Ramírez-Sánchez et al., 2020)—for each sentence pair. The document in Figure 1 illustrates examples of good, bad, and multi-way alignments along with their corresponding quality scores. We then map each alignment back to its specific source and target documents, maintaining the document-sentence relationships throughout the process. For any document that had multiple versions with the same URL, we explicitly

check that every sentence referenced in an alignment link actually exists in the final XML file to ensure it references the correct version(s). The output follows the standard `cesAlign` XML format[3], where each alignment links specific sentence IDs between source and target documents along with their quality scores.

MultiDocHPLT by pivoting via English As a "bonus" data release, English can be used as a pivot language to derive a corpus beyond the English-centric pairs. This enables the modeling and evaluation of DocMT between two non-English languages. The process is straightforward: if a document in a language and another document in another language are both aligned to the same English document, then we assume a direct alignment between the two documents. We pivot the sentence alignments in a similar way.

3.2 Data Statistics

In this section, we present the statistics of our English-centric dataset. We provide full tables in Appendix A: Table 7 details total documents and sentences per language, while Table 8 reports alignment statistics for each language pair. Specifically, for each language pair, we report the number of aligned document pairs (#doc pairs), the total number of alignments (#alignments), the average number of sentences per aligned document (avg #aligns./#docs), document length ratio calculated as the average number of sentences in English relative to the target language (avg #sent_en/#sent_xx), number of sentences per document (#sent/#docs), and the average alignment scores.

Across all language pairs, DocHPLT contains 87.8 million unique documents with 4.26 billion total sentences, averaging 48.6 sentences per document. The English collection dominates with 47.5 million documents (2.67 billion sentences), while individual non-English languages range from Japanese with 4 million documents (164 million sentences) down to Xhosa with 22 thousand documents (996 thousand sentences). These documents form 124 million aligned document pairs, with an average of 14.8 sentence-level alignments per document pair. Document coverage varies significantly: Japanese-English and Turkish-English each contribute over 11 million aligned document pairs, respectively, while under-represented languages like Sinhala-English (123 thousand document pairs), Uzbek-English (157 thousand document pairs), and Xhosa-English (44 thousand document pairs) have substantially smaller collections.

We observe considerable variations in document length ratios between aligned pairs, ranging from 3.91 (Malayalam-English), where the English documents are typically longer, to 0.84 (Arabic-English). Additionally, the average Bicleaner scores vary significantly, with language pairs like Arabic-English (0.700) demonstrating relatively high-quality alignments, whereas pairs such as Maltese-English (0.293) display substantially lower average alignment quality.

Alignment density Furthermore, we calculate alignment density (AD), which is defined as the proportion of aligned sentence pairs between two documents relative to the length of the longer document. Formally, given two documents D_{src} and D_{tgt}, with $|D_{\text{src}}|$ and $|D_{\text{tgt}}|$ denoting their respective sentence counts, the alignment density is computed as

$$AD = \frac{\text{\# of aligned sentence pairs}}{\max(|D_{\text{src}}|, |D_{\text{tgt}}|)} \quad (1)$$

Alignment density ranges between 0 (no aligned pairs) and 1 (perfect sentence-level coverage) if alignments are strictly one-to-one; however, since our alignment procedure allows one-to-many and many-to-one mappings, values above 1 are also possible. This feature may reveal the quality and the characteristics of the documents: an AD of exactly 1 could suggest that the documents were machine-translated (at the sentence level), whereas a very low AD might imply that they were accidentally matched, possibly due to high-frequency phrases or placeholders.

We observe considerable variation in AD across language pairs, e.g., Welsh-English (cy-en) and Afrikaans-English (af-en) show notably high average alignment densities (0.426 and 0.446, respectively), compared to languages like Farsi, Malayalam, and Marathi, where alignments are much sparser (0.153, 0.151, and 0.150, respectively). While some language pairs exhibit higher or lower densities, these scores are better understood relative to other scores rather than absolute terms.

We did not observe any consistent correlation between automatic quality metrics (BicleanerAI and CometKiwi) and AD values. Future work should investigate more carefully how AD should be interpreted and framed.

[3] https://opus.nlpl.eu/legacy/trac/wiki/DataFormats.html

4 Experiments and Findings

In order to test the usefulness of our dataset, we apply it to the task of fine-tuning LLMs for MT. Our first set of experiments tests different context lengths for this fine-tuning to see how much performance is affected by using the larger document contexts that DocHPLT enables. We then compare this best-performing fine-tuned configuration against off-the-shelf instruction-tuned models on the same test set. Finally, we investigate monolingual and multilingual fine-tuning for DocMT on DocHPLT. In the following sections, the notation of "src-trg" refers to the src-to-trg translation direction.

Languages We test translation from English into a total of 10 languages, chosen for their diversity in script, typology, and resource availability, as well as their inclusion in WMT24++ (Deutsch et al., 2025) (for testing) and CometKiwi (Rei et al., 2022) (for filtering). The languages are Arabic (ar), Catalan (ca), Hindi (hi), Estonian (et), Persian (fa), Finnish (fi), Icelandic (is), Korean (kr), Malayalam (ml), and Urdu (ur). It is worth noting that generating non-English is usually harder for LLMs compared to generating English.

Data processing We preprocess the documents for training by removing those with an AD below 0.3 or a document-averaged Bicleaner score below 0.3. We then discard unaligned segments in either source or target, and merge segments in a one-to-many alignment into a single segment. Finally, we filter data using CometKiwi (Rei et al., 2022) with SLIDE (Raunak et al., 2023): a window of 3 and a slide of 1. We retain document pairs with a CometKiwi score in the top 25$^{\text{th}}$ percentile for every language. This is to ensure that only high-quality parallel documents are used for training or evaluation.

Model training and inference We perform supervised fine-tuning (SFT) on Qwen2.5-7B-Instruct (Qwen et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with LoRA, rank 16 and alpha 32 (Hu et al., 2022), using the `open-instruct` toolkit[4]. Unless stated otherwise, our models are fine-tuned on 1000 documents per language, due to compute constraints. Our hyperparameters are listed in Appendix B. At test time, we always translate an entire source document in a single pass. LLM's chat template is always applied. All prompt information is detailed in Appendix C.

[4]https://github.com/allenai/open-instruct

	#test docs
en-fi	489
en-is	492
en-ko	497
en-ml	473
en-ur	486

Table 1: Test sizes after de-near-duplication (decontamination); always 500 for unseen language pairs.

Evaluation set We conduct evaluations on two test sets: a held-out set from DocHPLT and WMT24++, selected for their overlapping language coverage. We construct DocHPLT test by randomly sampling 500 documents per language from the CometKiwi-filtered corpora. We de-contaminate on the English side by computing Jaccard similarity over bigrams and removing any test document with a similarity above 0.8 to any training document. This ensures that our evaluation is not biased towards training on *similar* documents. The final test sizes are shown in Table 1.

Metrics We compute BLEU[5] (Papineni et al., 2002) and chrF++[6] (Popović, 2017) by treating each hypothesis document and reference document as a single string, and then averaging these scores across all documents. Our metric choice avoids the need for sentence-level alignment, which DocMT outputs do not guarantee. We note that while neural metrics such as COMET or LLM-as-a-judge are generally more reliable at the sentence level, their effectiveness in our document-level setting remains uncertain due to limited empirical validation, context support, and language coverage.

4.1 How much context do document-level models need?

Existing research on DocMT with LLM adopts distinct strategies to process data. Some approaches operate on a sentence level but use previous translations as context (Wu et al., 2024), some methods process fixed-size chunks (Alabi et al., 2025; Wicks et al., 2024; Post and Junczys-Dowmunt, 2024), translating each chunk separately, and some perform full document-to-document translation (Ramos et al., 2025). We start our experiments by training models using varying context lengths

[5]nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1
[6]nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1

		FT chunk (num sent)	DocHPLT					WMT24++				
			en-fi	en-is	en-ko	en-ml	en-ur	en-fi	en-is	en-ko	en-ml	en-ur
Qwen2.5-7B-Instruct	BLEU	1	8.39	20.13	10.39	**13.97**	11.05	11.49	**10.50**	5.68	**4.63**	**9.16**
		2	12.39	18.92	15.07	12.47	**11.48**	**12.81**	9.67	6.21	4.19	8.48
		5	12.80	28.99	22.69	10.20	8.94	12.38	9.55	**6.60**	2.93	5.72
		10	**13.87**	**32.54**	**23.71**	12.20	11.11	12.20	9.27	6.37	3.67	5.75
		doc2doc	8.35	24.65	15.57	9.35	5.05	9.40	6.02	6.02	1.21	2.44
	chrF++	1	30.51	40.49	23.84	**38.72**	32.76	36.37	**32.71**	19.36	**29.46**	**31.98**
		2	39.61	39.02	30.80	37.73	34.66	39.05	31.62	20.40	27.89	30.55
		5	42.12	50.00	39.24	33.05	30.71	**39.53**	31.30	**21.27**	22.90	25.49
		10	**44.01**	**53.74**	**40.87**	37.23	**34.72**	38.71	30.46	20.90	26.65	26.37
		doc2doc	35.12	46.10	30.70	32.60	24.05	34.25	24.59	19.30	16.67	16.42
Llama-3.1-8B-Instruct	BLEU	1	7.23	6.51	6.36	16.16	12.54	8.53	6.06	2.24	**4.32**	9.56
		2	10.49	11.53	10.98	**16.37**	14.11	11.25	7.62	2.85	3.86	9.39
		5	15.04	25.81	18.13	13.64	15.87	**12.76**	8.85	2.92	3.26	9.07
		10	**17.00**	**32.07**	**21.57**	15.94	**18.01**	12.74	**8.90**	3.16	4.00	**9.92**
		doc2doc	12.18	26.38	12.43	14.53	13.55	9.98	6.55	2.73	2.47	8.15
	chrF++	1	28.48	19.04	17.50	42.27	35.24	30.47	23.81	9.99	**28.04**	30.68
		2	34.89	30.65	25.17	43.07	37.84	35.09	26.58	12.11	26.45	30.70
		5	43.66	46.60	34.88	39.59	41.04	37.71	28.51	12.33	24.85	30.36
		10	**47.07**	**53.07**	**38.51**	**43.36**	**43.64**	**37.77**	**29.25**	12.81	27.15	**31.86**
		doc2doc	40.09	47.84	27.44	41.77	37.72	33.17	26.40	11.86	24.68	28.94

Table 2: Results from LLMs fine-tuned with different chunk sizes.

	Avg #tokens per doc	
	DocHPLT	**WMT24++**
en-ar	451	369
en-ca	550	402
en-hi	949	423
en-et	786	358
en-fa	582	397
en-fi	611	337
en-is	585	407
en-ko	602	338
en-ml	581	334
en-ur	822	448

Table 3: Average whitespace-delimited tokens per English document in DocHPLT and WMT24++ tests.

to find the best configuration.

Setup We build en-xx models for five target languages separately: Finnish, Icelandic, Korean, Malayalam, and Urdu. We fine-tune two open-source LLMs: Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. We fine-tune each model under five different context configurations: sentence-level (chunk 1, no context), chunks of 2, 5, and 10 sentences, as well as full document-to-document (doc2doc) training. For chunk-based training, we compute the loss only on target segments while providing source context. The total number of tokens is kept constant for all languages, despite the different data formats.

Results Our experiments in Table 2 reveal a clear and consistent pattern on the DocHPLT test set: measures of translation quality systematically improve as the input size increases from a single sentence to a 10-sentence chunk. As shown in the table, fine-tuning with 10-sentence chunks almost universally delivers the best performance across models, directions, and metrics. The gains are particularly dramatic for lower-resource pairs, such as en-is, where the BLEU score for Llama-3.1-8B-Instruct jumps from a sentence-level performance of 6.51 to 32.07. Nonetheless, full document-to-document training consistently underperforms the 10-sentence chunking strategy. This indicates that while substantial context is crucial, training LLMs on entire documents still poses challenges. This is consistent with Peng et al. (2025)'s findings that LLM-based translation degrades on longer documents.

However, we cannot observe a clear trend for WMT24++ regarding the training context size. The results are inconsistent, and the benefits of larger context windows are less clear. In several cases, smaller context windows or even simple sentence-level fine-tuning outperform the larger-context models, such as Qwen2.5-7B-Instruct on en-ml and en-ur. We hypothesize that the performance difference is due to document length variation as a domain bias. WMT24++ documents have roughly half the average number of tokens com-

			DocHPLT					WMT24++				
			en-fi	en-is	en-ko	en-ml	en-ur	en-fi	en-is	en-ko	en-ml	en-ur
Qwen2.5-7B-Instruct	BLEU	IT	11.01	10.42	14.33	3.05	3.79	11.57	6.12	5.90	2.42	3.97
		FT	**13.87**	**32.54**	**23.71**	**12.20**	**11.11**	**12.20**	**9.27**	**6.37**	**3.67**	**5.75**
	chrF++	IT	43.6	31.85	32.08	22.9	23.36	**40.76**	27.28	19.57	23.19	24.31
		FT	**44.01**	**53.74**	**40.87**	**37.23**	**34.72**	38.71	**30.46**	**20.9**	**26.65**	**26.37**
Llama-3.1-8B-Instruct	BLEU	IT	14.92	14.11	12.89	6.66	12.99	12.24	7.11	**3.78**	3.03	8.67
		FT	**17.00**	**32.07**	**21.57**	**15.94**	**18.01**	**12.74**	**8.90**	3.16	**4.00**	**9.92**
	chrF++	IT	46.42	38.45	30.67	32.59	39.40	**38.96**	27.88	**14.71**	25.32	30.71
		FT	**47.07**	**53.07**	**38.51**	**43.36**	**43.64**	37.77	**29.25**	12.81	**27.15**	**31.86**

Table 4: Results from prompting instruction-tuned (IT) LLMs and those further fine-tuned (FT) on DocHPLT.

pared to DocHPLT (Table 3), so most WMT24++ documents fit within a small chunk size. This creates a mismatch where training on larger chunk sizes is unnecessary or harmful, as longer contexts rarely occur in WMT24++.

These results show that the optimal context strategy for document-level translation is not absolute but is dependent on the test data characteristics. Based on our findings, we establish a training chunk size of 10 for all subsequent experiments.

4.2 Does fine-tuning on DocHPLT help document-level translation?

One key indicator of the usefulness of a data resource is whether practitioners can create better models using it. Although the origin of our data is web crawls, which may have been consumed by LLM pre-training, the parallelism signals are new in DocHPLT and not accessible through pre-training. Thus, in this section, we compare the results of fine-tuning LLMs to prompting baselines.

Setup We compare our fine-tuned models with the best-performing data configuration of chunk size 10 to their corresponding off-the-shelf instruction-tuned models. Evaluation is done on held-out DocHPLT test sets and WMT24++.

Results Table 4 shows that fine-tuning on DocHPLT produces notable improvements across nearly all settings, with gains inversely proportional to language resource levels. On DocHPLT test, lower-resourced languages see bigger jumps, e.g., in BLEU: 10.42 to 32.54 for Icelandic, 3.05 to 12.20 for Malayalam, and 3.79 to 11.11 for Urdu, whereas gains are more modest for higher-resourced languages, e.g., 11.01 to 13.87 for Finnish. On WMT24++, baseline prompting performance is generally poor, often below 6 BLEU, and improvements from fine-tuning persist but are smaller in absolute terms. This may be attributed to WMT24++'s domain mismatch (e.g., social and speech) with DocHPLT.

Our results suggest that off-the-shelf instruction-tuned models may already contain knowledge for these medium to low-resourced languages, yet fine-tuning on DocHPLT consistently improves performance across these languages. This underscores the value of our DocHPLT, which is the *first* to cater to those languages in this task. Nonetheless, we note that a higher performance does not necessarily indicate higher data quality—it may also be a result of greater exposure to a given language or to document-level input and output. A causal analysis will be useful, but for most of the languages we study, there is no suitable alternative data to compare to at the moment.

4.3 Does multilingual training improve over monolingual models?

While our monolingual fine-tuned models achieve significant gains over prompting baselines, deploying and maintaining separate models for each language presents scalability drawbacks. Furthermore, multilingual LLM fine-tuning may offer performance advantages over monolingual tuning (Chen et al., 2024). To test whether our multilingual DocHPLT can be exploited for cross-lingual transfer in training, in this section, we build and assess multilingual models. Particularly, we test on both seen and unseen languages to determine whether benefits extend beyond training languages.

Setup We compare three data configurations: a monolingual FT approach and two multilingual FT settings, resulting in three models:

- Mono_{1K}: a monolingual FT data approach that uses 1000 documents per language.

	Seen Languages		DocHPLT					WMT24++				
			en-fi	en-is	en-ko	en-ml	en-ur	en-fi	en-is	en-ko	en-ml	en-ur
Qwen2.5-7B-Instruct	BLEU	Mono_{1K}	13.87	32.54	**23.71**	12.20	11.11	12.20	9.27	6.37	3.67	5.75
		Multi_{1K}	10.31	24.11	20.12	5.07	4.35	11.66	6.62	**6.76**	1.43	3.45
		Multi_{5K}	**14.05**	**35.13**	23.60	**14.70**	**13.07**	**13.42**	**10.02**	6.62	**4.18**	**7.70**
	chrF++	Mono_{1K}	**44.01**	53.74	**40.87**	37.23	34.72	38.71	30.46	20.90	26.65	26.37
		Multi_{1K}	38.01	44.70	37.56	26.07	22.32	37.56	26.40	**21.50**	18.44	21.12
		Multi_{5K}	**44.09**	**56.74**	40.56	**40.28**	**37.16**	40.35	**32.24**	21.23	**27.28**	**29.57**
Llama-3.1-8B-Instruct	BLEU	Mono_{1K}	**17.00**	32.07	**21.57**	15.94	**18.01**	12.74	**8.90**	3.16	**4.00**	9.92
		Multi_{1K}	13.57	25.75	17.58	10.15	13.24	11.23	7.03	3.24	2.83	7.62
		Multi_{5K}	16.57	**34.21**	21.04	**17.01**	17.55	**13.39**	8.46	3.33	3.87	**10.13**
	chrF++	Mono_{1K}	**47.07**	53.07	**38.51**	43.36	**43.64**	37.77	**29.25**	12.81	**27.15**	**31.86**
		Multi_{1K}	43.04	47.64	34.74	36.55	37.88	36.07	26.04	12.63	24.33	27.31
		Multi_{5K}	45.00	**55.39**	37.83	**44.69**	42.73	38.15	28.47	12.53	26.73	31.77

Table 5: Results from monolingual and multilingual fine-tuning for *seen languages*.

	Unseen Languages		DocHPLT					WMT24++				
			en-et	en-ca	en-hi	en-fa	en-ar	en-et	en-ca	en-hi	en-fa	en-ar
Qwen2.5-7B-Instruct	BLEU	IT	**7.39**	26.47	**11.40**	**8.34**	13.63	**7.82**	**19.41**	**9.87**	**10.14**	8.65
		Multi_{1K}	4.96	**26.59**	8.22	7.28	**15.85**	6.96	18.78	7.44	8.80	**10.09**
		Multi_{5K}	4.74	25.41	7.01	4.21	14.28	6.48	18.06	7.43	4.96	9.74
	chrF++	IT	**36.34**	**55.59**	**35.46**	**36.12**	39.44	**33.04**	**47.19**	**33.30**	**36.16**	31.84
		Multi_{1K}	26.85	54.92	27.74	31.83	**42.15**	28.02	45.17	26.47	31.67	**34.04**
		Multi_{5K}	27.28	53.81	24.99	23.84	39.12	27.62	44.72	27.22	24.45	34.02
Llama-3.1-8B-Instruct	BLEU	IT	**11.50**	**32.19**	22.13	**13.26**	**12.62**	**9.20**	**20.82**	12.58	9.50	**6.94**
		Multi_{1K}	8.95	31.45	23.08	12.65	11.33	8.29	19.60	12.66	**9.60**	6.37
		Multi_{5K}	8.73	30.17	**23.95**	11.87	10.55	7.61	19.29	**13.40**	9.34	6.30
	chrF++	IT	**41.88**	**57.73**	47.56	**40.80**	**37.51**	**32.87**	**44.46**	**35.44**	**32.68**	**28.26**
		Multi_{1K}	35.41	56.75	47.68	38.79	33.82	29.47	42.12	34.68	31.84	25.98
		Multi_{5K}	34.08	55.63	**47.96**	37.76	32.44	28.05	41.83	35.35	31.04	25.19

Table 6: Results from prompting instruction-tuned (IT) LLMs and multilingual fine-tuning for *unseen languages*.

- Multi_{1K}: a multilingual setting that uses 1000 documents combined, with 200 from each of the 5 languages, intended to match the total size for monolingual FT.

- Multi_{5K}: another multilingual setting that uses 5000 documents in total, with 1000 from each language, intended to match the size for each language in monolingual FT.

All models are trained with consistent hyperparameters as in Appendix B. We stick to our best-performing data configuration of chunk size 10.

We evaluate those models on DocHPLT and WMT24++ for all 5 training languages and 5 additional unseen languages: Arabic (ar), Catalan (ca), Estonian (et), Hindi (hi), and Persian (fa). These unseen languages are selected for their linguistic and/or script relation with the training languages.

Results Table 5 compares multilingual to monolingual FT, showing that multilingual advantages are model-dependent. For Qwen2.5-7B-Instruct, Multi_{5K} outperforms Mono_{1K} and Multi_{1K} consistently, whereas Llama-3.1-8B-Instruct displays a mixed pattern. Taking a closer look at the languages, Icelandic and Malayalam always improve with multilingual training, regardless of the LLM. In Table 6, for unseen languages, we see that the off-the-shelf IT models are usually better than multilingual fine-tuning.

In general, multilingual fine-tuning produces inconsistent results: it improves DocMT performance over monolingual fine-tuning for some LLMs, but we find almost no zero-shot cross-lingual transfer. Our observations from a small-scale multilingual experiment warrant further investigation by scaling the model choices and sizes,

as well as the number of languages which is supported by DocHPLT.

5 Conclusion

We introduced a pipeline to derive a document-level corpus with rich metadata and presented the outcome, DocHPLT, the largest publicly available document-level translation dataset with 124 million aligned document pairs across 50 languages paired with English. The utility of our massively multilingual dataset has been demonstrated through experiments: fine-tuning LLMs on our data improved over prompting baselines, and multilingual training surpassed monolingual models, though zero-shot transfer to unseen languages remained challenging. Our experiments also revealed challenges in DocMT: full document-to-document training and generalization to other document domains. Future work may use DocHPLT data for further investigations such as large-scale training, data filtering, data synthesis, and DocMT metric study.

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546].

Dayyán O'Brien is also supported by a G-Research NextGen Scholarship, part of the UKRI AI Centre for Doctoral Training in Responsible and Trustworthy in-the-world Natural Language Processing (grant ref: EP/Y030656/1).

We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Regular Access call.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*.

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Jesujoba O Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, and others. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. *arXiv preprint arXiv:2501.06374*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, and others. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Laurie Burchell, Ona De Gibert Bonet, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, and others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, and others. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024*

Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and others. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. Source-primed multi-turn conversation helps large language models translate documents. *arXiv preprint arXiv:2503.10494*.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2022. A bilingual parallel corpus with discourse annotations. *arXiv preprint arXiv:2210.14667*.

Linghao Jin, Li An, and Xuezhe Ma. 2024. Towards chapter-to-chapter context-aware literary translation via large language models. *arXiv preprint arXiv:2407.08978*.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq Joty, and Dominika Basaj. 2020. Can your context-aware MT system pass the DiP benchmark tests?: Evaluation benchmarks for discourse phenomena in machine translation. *arXiv preprint arXiv:2004.14607*.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys*.

Marcin Miłkowski and Jarosław Lipski. 2011. Using SRX standard for sentence segmentation. In *Human Language Technology. Challenges for Computer Science and Linguistics*.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*.

Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2023. Document-level language models for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*.

Matt Post and Marcin Junczys-Dowmunt. 2024. Evaluation and large-scale training for contextual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*.

Qwen, Baosong Yang An Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, and others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2505.09388*.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and André FT Martins. 2025. Multilingual contextualization of large language models for document-level machine translation. *arXiv preprint arXiv:2504.12140*.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Evaluating metrics for document-context evaluation in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and others. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Michelle Wastl, Jannis Vamvas, Selena Calleri, and Rico Sennrich. 2025. 20min-XD: A comparable corpus of Swiss news articles. In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. Recovering document annotations for sentence-level bitext. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *arXiv preprint arXiv:2401.06468*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

A Dataset Statistics

	#sentences	#docs	#deduped docs
af	16,416,841	297,636	286,861
ar	65,482,300	2,271,167	2,196,334
az	12,202,189	332,742	321,500
be	10,672,952	212,121	203,758
bg	80,018,549	1,746,301	1,669,696
bn	10,473,372	414,099	405,339
bs	20,635,243	514,615	488,093
ca	47,905,003	1,198,217	1,131,468
cy	8,908,119	265,261	253,040
en	2,665,945,834	47,484,349	45,995,228
eo	6,115,355	119,196	103,858
et	33,684,509	774,561	747,075
eu	6,783,654	189,347	175,318
fa	24,837,952	810,029	785,963
fi	111,615,913	2,445,791	2,341,993
ga	6,398,081	172,167	166,516
gl	10,657,570	233,545	215,009
gu	3,202,679	108,507	106,423
he	38,077,820	1,190,198	1,149,349
hi	37,592,475	1,336,090	1,315,174
hr	52,267,826	1,063,347	1,009,227
is	12,571,982	274,078	265,088
ja	164,136,152	4,032,689	3,934,457
kk	5,948,866	140,082	135,689
kn	4,463,262	123,053	120,996
ko	84,527,642	2,058,811	2,003,338
lt	48,692,264	1,031,628	995,288
lv	37,426,957	796,659	766,138
mk	12,465,228	307,055	292,992
ml	2,925,457	115,189	111,721
mr	3,066,703	128,808	126,552
ms	51,150,528	978,185	942,418
mt	6,328,544	141,088	137,104
nb	89,189,502	1,884,362	1,809,266
ne	1,549,852	74,579	73,691
nn	4,228,079	93,285	78,426
si	1,497,375	50,605	48,730
sk	70,057,465	1,461,804	1,406,726
sl	37,501,647	797,858	765,171
sq	11,475,561	328,651	317,315
sr	21,620,629	407,440	386,829
sw	8,409,824	185,287	178,185
ta	6,790,864	215,564	208,573
te	5,131,680	141,279	138,727
th	16,134,265	676,699	655,628
tr	100,380,235	3,884,137	3,767,266
uk	89,841,883	1,955,041	1,891,287
ur	5,479,098	234,708	228,952
uz	3,502,356	69,440	68,191
vi	87,511,126	1,986,258	1,940,095
xh	995,556	21,561	20,797
total	**4,264,894,818**	**87,775,169**	**84,882,858**

Table 7: A summary of DocHPLT documents and sentences per language.

	#doc pairs	#alignments	avg #aligns. /#doc	avg #sents_en /#sents_xx	#sents/#docs		avg BLEUalign	avg Bicleaner	avg align. density
					en	xx			
af-en	1,121,166	29,496,715	26.3	1.38	85.1	108.5	0.551	0.418	0.446
ar-en	4,405,876	54,747,241	12.4	0.84	35.7	56.6	0.468	0.700	0.280
az-en	732,657	10,289,514	14.0	1.26	53.1	64.7	0.443	0.482	0.334
be-en	709,129	14,728,785	20.8	1.37	87.7	104.9	0.543	0.556	0.324
bg-en	6,016,906	93,051,525	15.5	1.34	65.9	79.9	0.541	0.582	0.285
bn-en	1,039,423	7,851,362	7.6	0.89	34.4	69.8	0.446	0.577	0.182
bs-en	1,443,819	17,704,604	12.3	1.20	65.4	95.6	0.512	0.516	0.268
ca-en	3,582,267	63,520,169	17.7	1.20	60.4	87.6	0.562	0.620	0.335
cy-en	721,671	12,632,309	17.5	1.15	45.4	60.9	0.577	0.618	0.426
eo-en	482,452	8,677,590	18.0	3.61	147.6	82.1	0.511	0.474	0.246
et-en	2,484,493	40,019,712	16.1	1.96	74.5	56.7	0.502	0.501	0.311
eu-en	616,924	8,245,785	13.4	2.85	88.4	52.0	0.493	0.402	0.294
fa-en	1,880,900	11,884,837	6.3	2.69	76.2	40.1	0.423	0.544	0.153
fi-en	8,532,601	135,452,163	15.9	1.80	76.0	61.9	0.546	0.555	0.307
ga-en	557,716	10,060,287	18.0	1.85	61.2	49.6	0.613	0.488	0.419
gl-en	988,176	15,903,011	16.1	3.06	120.3	69.4	0.533	0.532	0.256
gu-en	306,386	3,358,243	11.0	2.65	91.6	52.7	0.476	0.500	0.187
he-en	4,190,235	49,247,941	11.8	2.91	85.7	40.8	0.537	0.577	0.220
hi-en	3,502,520	32,907,313	9.4	2.55	60.9	36.5	0.479	0.609	0.196
hr-en	3,574,689	54,626,216	15.3	1.77	82.0	72.3	0.537	0.528	0.302
is-en	1,097,797	19,959,668	18.2	2.02	77.4	52.6	0.498	0.474	0.333
ja-en	11,828,819	144,978,567	12.3	1.75	63.7	49.9	0.462	0.382	0.181
kk-en	243,579	4,197,879	17.2	1.47	72.9	60.7	0.446	0.559	0.381
kn-en	355,117	5,270,814	14.8	2.65	124.1	71.2	0.446	0.515	0.200
ko-en	6,479,547	106,313,693	16.4	1.98	78.9	54.7	0.526	0.572	0.237
lt-en	3,948,829	62,315,769	15.8	1.78	74.5	64.5	0.538	0.546	0.283
lv-en	3,104,028	53,107,619	17.1	1.96	77.8	63.9	0.555	0.573	0.308
mk-en	961,749	17,710,874	18.4	2.03	94.8	65.8	0.518	0.576	0.319
ml-en	298,334	2,211,378	7.4	3.91	89.6	36.9	0.427	0.459	0.151
mr-en	372,093	2,567,437	6.9	3.44	70.4	33.4	0.432	0.446	0.150
ms-en	3,887,463	69,632,512	17.9	2.01	82.2	65.6	0.551	0.390	0.289
mt-en	477,497	9,464,200	19.8	1.72	69.7	59.7	0.605	0.293	0.407
nb-en	6,596,166	105,226,440	16.0	1.61	66.7	58.8	0.542	0.556	0.308
ne-en	201,928	1,415,859	7.0	3.03	57.8	26.1	0.448	0.394	0.169
nn-en	413,279	4,396,370	10.6	3.56	113.0	55.9	0.445	0.421	0.164
si-en	123,803	1,338,609	10.8	2.36	83.5	51.9	0.474	0.442	0.219
sk-en	5,262,604	81,849,513	15.6	1.56	73.6	66.0	0.536	0.597	0.293
sl-en	2,334,208	41,082,011	17.6	1.73	80.2	68.7	0.503	0.536	0.329
sq-en	910,599	15,055,014	16.5	1.93	78.0	58.6	0.529	0.515	0.382
sr-en	1,307,126	25,315,953	19.4	1.88	106.8	78.4	0.541	0.492	0.335
sw-en	581,466	12,214,107	21.0	1.95	98.2	84.0	0.557	0.340	0.348
ta-en	583,034	5,804,724	10.0	2.68	84.5	41.8	0.458	0.434	0.190
te-en	389,858	5,202,332	13.3	2.83	123.4	68.4	0.466	0.495	0.178
th-en	2,438,548	18,656,911	7.7	2.76	57.5	30.5	0.501	0.531	0.197
tr-en	11,815,778	120,528,089	10.2	2.80	62.4	34.5	0.520	0.503	0.215
uk-en	5,364,321	88,197,354	16.4	1.64	80.8	68.5	0.516	0.608	0.312
ur-en	618,996	5,471,488	8.8	2.94	70.2	39.1	0.463	0.508	0.198
uz-en	156,796	3,300,674	21.1	1.61	85.2	76.7	0.461	0.492	0.369
vi-en	5,089,734	66,322,073	13.0	1.72	76.2	61.2	0.413	0.626	0.235
xh-en	44,001	1,276,014	29.0	1.67	96.3	101.3	0.477	0.443	0.407
Average	2,483,542	35,495,785	14.8	2.11	79.4	61.8	0.503	0.510	0.277
Total	124,177,103	1,774,789,267							

Table 8: A summary of DocHPLT alignment statistics by language pair.

B Hyperparamters

Below, we list the hyperparameters used during training.

Parameter	Value
Learning Rate	5e-04
LR Scheduler Type	Linear
Warmup Ratio	0.1
Weight Decay	0.0
Per Device Train Batch Size	2
Gradient Accumulation Steps	4
Number of Train Epochs	1
LoRA Rank	16
LoRA Alpha	32
Seed	1729

Table 9: Training Hyperparameters

C Prompts

C.1 Overview

We use the same prompt for SFT and during inference for both off-the-shelf instruction-tuned and fine-tuned models. LLM's chat template is always applied.

We illustrate chunk-based translation and full document-to-document translation using the task of English to Catalan translation.

C.2 Chunk-based translation

Template (chunk size 2):

> Translate the following source segment from [SOURCE LANGUAGE] into [TARGET LANGUAGE].
>
> [SOURCE LANGUAGE]: [SOURCE TEXT]
>
> [TARGET LANGUAGE]: [TARGET TEXT]

Example:

> Translate the following source segment from English into Catalan.
>
> English: Online workshops are organized every month. The results will be shared with the community.
>
> Catalan: Cada mes s'organitzen tallers en línia. Els resultats es compartiran amb la comunitat.

C.3 Document-to-document translation

Template

> Translate the following source document from [SOURCE LANGUAGE] into [TARGET LANGUAGE].
>
> [SOURCE LANGUAGE]: [SOURCE DOCUMENT]
>
> [TARGET LANGUAGE]: [TARGET DOCUMENT]

Example:

> Translate the following source document from English into Catalan.
>
> English: Our proposals with you in mind. We suggest.... Castelló d'Empúries is situated in the heart of the Aiguamolls Natural Park. Stay at a house in the historic center of Castelló d' Empúries Check opening times and escape from the hustle and bustle of the city with a visit you will love. A weekend to explore Empordà by bike. Here you will also find events, fairs and festivals that are held close to Hostal Casa Clara.
>
> Catalan: Les nostres propostes pensades per a vosaltres. Us suggerim... Castelló d'Empúries està situat al bell mig del Parc Natural dels Aiguamolls. Les teves vacances en una casa al centre històric de Castelló d'Empúries Consulta els horaris i fes una visita que t'encantarà i et farà desconnectar del brogit de ciutat. Un cap de setmana en bicicleta per conèixer l'Empordà. Aquí també hi trobaràs esdeveniments, fires i festes populars que es fan prop de l'Hostal Casa Clara

SONAR-SLT: Multilingual Sign Language Translation via Language-Agnostic Sentence Embedding Supervision

Yasser Hamidullah[1] **Shakib Yazdani**[1] **Cennet Oguz**[1]
Josef van Genabith[1] **Cristina España-Bonet**[1,2]

[1]German Research Center for Artificial Intelligence (DFKI GmbH),
Saarland Informatics Campus, Saarbrücken, Germany
[2]Barcelona Supercomputing Center (BSC-CNS), Barcelona, Catalonia, Spain

{yasser.hamidullah,shakib.yazdani,cennet.oguz,josef.van_genabith,cristinae}@dfki.de

Abstract

Sign language translation (SLT) is typically trained with text in a single spoken language, which limits scalability and cross-language generalization. Earlier approaches have replaced gloss supervision with text-based sentence embeddings, but up to now, these remain tied to a specific language and modality. In contrast, here we employ language-agnostic, multimodal embeddings trained on text and speech from multiple languages to supervise SLT, enabling direct multilingual translation. To address data scarcity, we propose a coupled augmentation method that combines multilingual target augmentations (i.e. translations into many languages) with video-level perturbations, improving model robustness. Experiments show consistent BLEURT gains over text-only sentence embedding supervision, with larger improvements in low-resource settings. Our results demonstrate that language-agnostic embedding supervision, combined with coupled augmentation, provides a scalable and semantically robust alternative to traditional SLT training.[1]

1 Introduction

Sign languages (SLs) are inherently visual and culturally embedded. Each SL has evolved independently and is closely tied to the communities and spoken languages of its region. As a result, most sign language translation (SLT) datasets are built around a *single* sign–spoken language pair (e.g., DGS→German), which makes it difficult to scale models across languages or to combine datasets. Training a system for a new target language typically requires a separate model and fresh parallel data collection.

Historically, SLT systems have relied on manually provided *gloss supervision* (Camgoz et al., 2018), discrete word-like labels whose design and availability are language-, culture-, and region-specific. Even *gloss-free* SLT approaches assume that sign inputs should be supervised by text from the co-occurring spoken language, keeping the learning signal tied to a single language (Gong et al., 2024; Wong et al., 2024; Chen et al., 2024; Hamidullah et al., 2022) and limiting cross-dataset reuse and generalization.

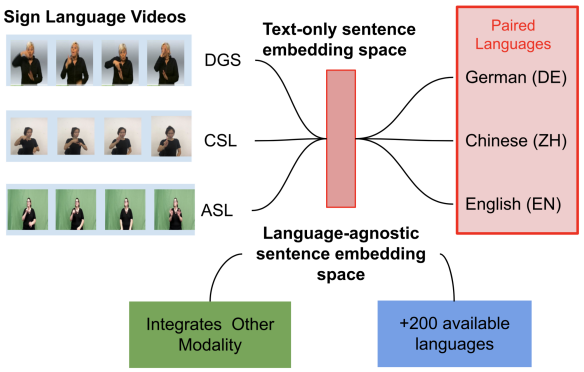


Figure 1: Text-only vs. language-agnostic sentence embedding supervision.

Recent work by Hamidullah et al. (2024) has reduced the reliance on glosses by supervising SLT with *text-based sentence embeddings*. This yields better semantic alignment, but the embeddings remain modality-specific and typically require dataset-specific fine-tuning. Furthermore, compared to large pre-trained models that exploit vast text corpora, these text-only embeddings show limited cross-lingual transfer and reduced robustness. This raises the key question: *Can language-agnostic, multimodal sentence embedding supervision replace text-only alignment in SLT?* We hypothesize that **language-agnostic, multimodal sentence embeddings** can reduce the residual dependence on text. Concretely, we build on SONAR (Duquenne et al., 2023), a pretrained multilingual and multimodal embedding space that jointly rep-

[1]We release the code, models, and features to facilitate further research. Github repository: https://github.com/DFKI-SignLanguage/sonar-slt.git; Huggingface: https://huggingface.co/mtmlt

Proceedings of the Tenth Conference on Machine Translation, pages 301–313
November 8-9, 2025

resents text and speech. SONAR embeddings are claimed to be language-agnostic. Our approach aligns sign representations directly with language-agnostic semantic vectors, thereby decoupling supervision from any specific spoken language and removing the need for glosses. Our model integrates multiple modalities and supports direct supervision across all 200 languages covered by SONAR (see Figure 1). In contrast to prior systems that relied on additional stages or separate models for multi-target translation, our method enables *direct* translation into multiple languages within a single model.

A major obstacle for SLT is the scarcity of annotated data. Recent work on self-supervised pre-training from unannotated or anonymized data (Rust et al., 2024) has shown promise in addressing this challenge. This motivates our second question: *Can target-language augmentation further alleviate data scarcity and enhance robustness, particularly when combined with video augmentation?*

Our **coupled multiple target language and video perturbation augmentation** strategy addresses these challenges by combining (i) *target-language augmentation*, which pairs each sign sample with parallel sentences in multiple languages, and (ii) *video augmentation*, which perturbs the visual stream through spatial, temporal, and photometric transformations. These augmentations are complementary: multiple target-language augmentation strengthens semantic supervision without requiring new sign recordings, while video augmentation improves the invariance of the sign encoder. Together, they yield a more robust SLT model and provide a scalable, semantically grounded alternative to traditional training, unifying supervision across languages and modalities while reducing dependence on language-, culture-, and region-specific annotations. In all, our contributions can be summarized as:

- **Language-agnostic supervision.** We align signs to a multilingual, multimodal embedding space, removing reliance on language-specific text or glosses.
- **Coupled augmentation.** We jointly apply multilingual target augmentation and video perturbations to improve robustness and reduce data scarcity.
- **Direct multilingual decoding.** Our model translates into multiple spoken languages in a single step, without pivots or extra fine-tuning.
- **Open-source resources.** We release a Hugging Face–compatible *visual extension of* SONAR and model port to enable reproducibility and further work.

2 Related Work

2.1 Sign Language Representation

Traditional SLT systems rely on glosses —textual labels that represent signs— as an intermediate representation. MSKA-SLT (Guan et al., 2025) remains a strong baseline using glosses, reporting ∼29 BLEU on PHOENIX-2014T (Camgoz et al., 2018). However, glosses are neither universal nor standardized: they are tightly coupled to specific languages, cultures, and regions. Moreover, producing gloss annotations is highly time-consuming, requiring expert linguistic knowledge (Müller et al., 2023b).

In parallel, gloss-free SLT has emerged, enabling training on weakly annotated datasets exceeding 1,000 hours for some sign languages (Uthus et al., 2023).[2] Hamidullah et al. (2024) aligns sign language videos with sentence-level text embeddings. This supervision avoids feeding long, fine-grained frame sequences to the decoder, thereby reducing redundancy in video features, lowering the need for aggressive masking, and encouraging learning at the sentence-semantic level. While intermediate supervision of visual blocks is common in multimodal models, compressing video into a sentence-level embedding before decoding improves semantic grounding and flexibility in target text generation. Nevertheless, current approaches (Hamidullah et al., 2024; Gueuwou et al., 2025b) remain limited by their reliance on *text-only* embedding spaces with restricted language coverage, constraining augmentation and cross-sign transfer.

2.2 Large Language Models in SLT

Complementary approaches leverage large language models (LLMs). SignLLM (Gong et al., 2024) discretizes videos into tokens and prompts a frozen LLM; Sign2GPT (Wong et al., 2024) feeds pseudo-glosses to XGLM, reporting ∼22 BLEU on

[2] A *weakly annotated dataset* provides only coarse or noisy supervision. For instance, YouTube-ASL datasets are collected from online videos where annotations rely solely on automatically generated or the provided subtitles, without manual realignment, leading to potential inaccuracies and temporal misalignments.

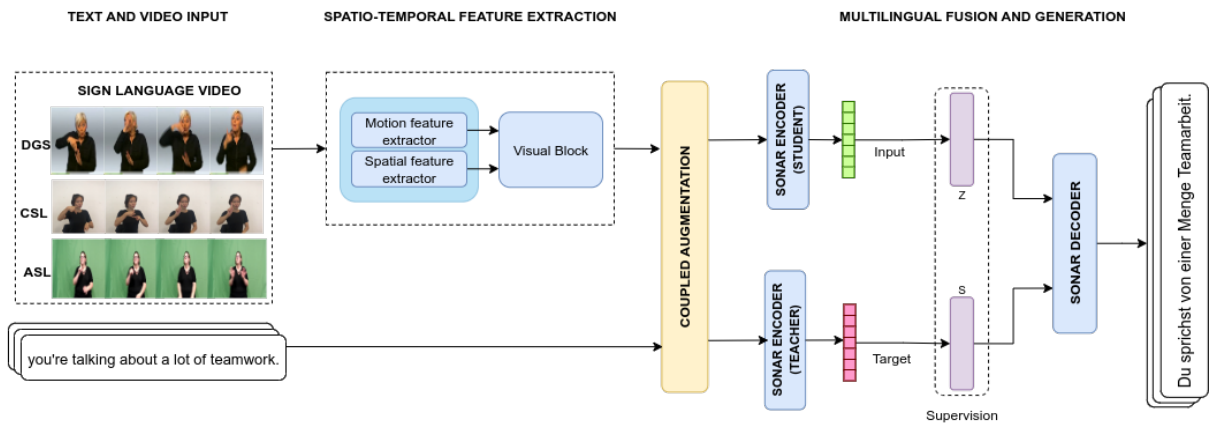


Figure 2: Overall architecture of our SONAR-SLT model. Visual inputs are processed through spatial and spatio-temporal encoders, fused using (Hwang et al., 2025) and encoded into a semantic vector aligned with multilingual sentence embeddings.

PHOENIX-2014T and ∼15 BLEU on CSL-Daily. SpaMo (Hwang et al., 2025) employs a straightforward approach that extracts spatial and motion features from sign language videos and utilizes a low-rank adapter to fine-tune an LLM for sign language translation. Chen et al. (2024) introduced FLa-LLM, a two-stage, gloss-free framework that first pre-trains the visual encoder and then fine-tunes a pre-trained LLM for the downstream SLT task. These methods inherit LLM fluency but are largely monolingual and require substantial tokenization and training overhead. In contrast, our PEFT-based SONAR adapters maintain multilinguality without retraining a large decoder on discretized video tokens. More recent work has explored large-scale pre-training to improve sign language understanding, with Uni-Sign (Li et al., 2025) proposing a unified generative framework that treats downstream tasks as SLT and incorporates prior-guided fusion.

2.3 Multilingual SLT Datasets and Models

Despite these advances, large-scale multilingual datasets (Uthus et al., 2023; Yazdani et al., 2025b) remain scarce and noisy. Crawled web data increases coverage but introduces label and alignment errors that current models struggle to absorb, leading many studies to focus on a single language or a small set of cleaner corpora. Additionally, performance often varies widely even within the same language due to differences in feature pipelines and recording conditions.

Multilingual SLT models also remain in their early stages. MLSLT (Yin et al., 2022) covers ten European sign languages via a routing mechanism, while JWSign (Gueuwou et al., 2023) scales to 98 languages with language-ID tokens. More recently, Sign2(LID+Text) (Tan et al., 2025) incorporated token-level language identification with a CTC loss, achieving competitive results. In addition, Yazdani et al. (2025a) explored continual learning for multilingual SLT. Recent work applies heavy pre-processing (Gueuwou et al., 2025b,a), sometimes obscuring whether improvements arise from better SLT modeling or dataset-specific engineering. Both gloss-based and gloss-free methods perform best when signer distance, camera setup, and motion characteristics closely match training conditions.

3 Methodology

3.1 System Overview

We propose **SONAR-SLT**, a modular SLT framework that decouples *semantic understanding* from *text generation*. As illustrated in Figure 2, the system first maps an input sign language video into a multilingual, multimodal semantic space, and then (optionally) decodes from this space into a chosen spoken language. This design allows training on heterogeneous sign language datasets, supports multilingual supervision, and removes the need for gloss annotations. A detailed architecture is presented in Appendix A.1 and summarized in the next subsections.

3.2 Visual Feature Extraction and Encoding

The first stage maps raw video frames into a compact visual embedding. Let $x = (f_1, \ldots, f_T)$ denote a sign language video of T frames. We extract per-frame spatial features $\mathbf{s}_t$ with ViT (Dosovitskiy

et al., 2020) and spatio-temporal motion features $\mathbf{m}_t$ with VideoMAE (Tong et al., 2022). These are fused through a lightweight block (1D Conv followed by a multi-layer perceptron) $\mathcal{F}$ (Hwang et al., 2025):

$$\mathbf{h}_t = \mathcal{F}(\mathbf{s}_t, \mathbf{m}_t), \qquad t = 1, \ldots, T. \tag{1}$$

A Transformer-based encoder $\mathcal{E}_v$ contextualizes the sequence:

$$\mathbf{z}_{1:T} = \mathcal{E}_v(\mathbf{h}_{1:T}). \tag{2}$$

Finally, temporal pooling (mean or attention) produces a global visual embedding $\mathbf{z} \in \mathbb{R}^d$:

$$\mathbf{z} = \mathrm{Pool}(\mathbf{z}_{1:T}). \tag{3}$$

3.3 Semantic Alignment

Next, we align sign-derived embeddings with multilingual textual embeddings. We adopt a pretrained multilingual, multimodal sentence encoder $\mathcal{E}$ (i.e., SONAR). Given a reference sentence y, we obtain its semantic embedding:

$$\mathbf{s} = \mathcal{E}_{txt}(y) \in \mathbb{R}^d. \tag{4}$$

The visual encoder is trained to align $\mathbf{z}$ with $\mathbf{s}$. Alignment can be done via a squared ℓ_2 loss as per (Duquenne et al., 2023; Hamidullah et al., 2024):

$$\mathcal{L}_{\text{sem}} = \|\mathbf{z} - \mathbf{s}\|_2^2. \tag{5}$$

We also consider a cosine similarity loss,

$$\mathcal{L}_{\cos} = 1 - \frac{\langle \mathbf{z}, \mathbf{s} \rangle}{\|\mathbf{z}\|_2 \, \|\mathbf{s}\|_2}, \tag{6}$$

used either alone ($\mathcal{L}_{\text{sem}} = \mathcal{L}_{\cos}$) or combined with the MSE above:

$$\mathcal{L}_{\text{sem}} = \alpha \|\mathbf{z} - \mathbf{s}\|_2^2 + \beta \mathcal{L}_{\cos}, \qquad \alpha, \beta \geq 0. \tag{7}$$

Target-language augmentation. To enforce language-agnostic supervision, each reference sentence is paired with K translations $\{y^{(k)}\}_{k=1}^K$ (from the embedding decoder). At each iteration, one translation $y^{(k)}$ is sampled and encoded as $\mathbf{s} = \mathcal{E}_{txt}(y^{(k)})$.

3.4 Multilingual Generation from the Semantic Vector

We then decode into natural language from the semantic embedding. A pretrained decoder $\mathcal{D}$ from SONAR generates text from a semantic vector and a target language token ℓ. Conditioned on the sign-derived and semantically text-aligned (Section 3.3) embedding $\mathbf{z}$, the decoder is trained with teacher forcing:

$$\mathcal{L}_{\text{ce}} = -\sum_{t=1}^{T_y} \log p_\theta(y_t \mid y_{<t}, \mathbf{z}, \ell). \tag{8}$$

3.5 Auto-Encoding (Decoder Anchoring)

To keep the decoder aligned to the pretrained semantic space, we introduce an auto-encoding step. Specifically, the decoder reconstructs the target sentence directly from its text-derived embedding $\mathbf{s}$:

$$\mathcal{L}_{\text{ae}} = -\sum_{t=1}^{T_y} \log p_\theta(y_t \mid y_{<t}, \mathbf{s}, \ell). \tag{9}$$

This mirrors SONAR's original training and prevents drift, while the visual encoder learns to project videos into the same space.

3.6 Optional Contrastive Alignment

We optionally strengthen alignment through a symmetric InfoNCE loss (van den Oord et al., 2018). For a batch $\{(\mathbf{z}_i, \mathbf{s}_i)\}_{i=1}^N$, we define similarity as $\mathrm{sim}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\tau}$ (temperature $\tau > 0$, optional ℓ_2 normalization). The corresponding loss is:

$$\mathcal{L}_{\text{nce}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{s}}_i / \tau)}{\sum_{j=1}^N \exp(\hat{\mathbf{z}}_i^\top \hat{\mathbf{s}}_j / \tau)} - \log \frac{\exp(\hat{\mathbf{s}}_i^\top \hat{\mathbf{z}}_i / \tau)}{\sum_{j=1}^N \exp(\hat{\mathbf{s}}_i^\top \hat{\mathbf{z}}_j / \tau)} \right]. \tag{10}$$

3.7 Joint Training Objective

The final loss combines all components:

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{sem}} \mathcal{L}_{\text{sem}} + \lambda_{\text{ce}} \mathcal{L}_{\text{ce}} + \lambda_{\text{ae}} \mathcal{L}_{\text{ae}} + \lambda_{\text{nce}} \mathcal{L}_{\text{nce}}, \tag{11}$$

with non-negative weights λ_i (setting $\lambda_{\text{nce}} = 0$ disables the contrastive term).

3.8 Cross-Lingual and Multi-Sign Dataset Fusion

Finally, we leverage the language-agnostic semantic space for dataset fusion. Because supervision is defined independently of any specific spoken language, videos from different sign languages can be trained jointly with textual supervision in *any* available language. For example, German sign language

Dataset	Language	Domain	#Videos	#Sent.	Vocab.	Split (train/dev/test)
PHOENIX-2014T	DGS → German	Weather Forecast	∼7k	∼8k	∼3k	7,096 / 519 / 642
CSL-Daily	CSL → Chinese	Daily Communication	∼20k	∼25k	∼5k	18,401 / 1,078 / 1,057

Table 1: Characteristics of the datasets used in our experiments.

videos annotated in German can be re-aligned with English, French, or Chinese translations via $\mathcal{E}$, allowing unified training across datasets such as PHOENIX-2014T and CSL-Daily. This enables direct multi-target translation without glosses and facilitates fusion of heterogeneous sign-language corpora.

4 Experiments

4.1 Datasets

We evaluate our approach on the following datasets:

- **PHOENIX-2014T** (Camgoz et al., 2018): German Sign Language (DGS) weather forecast videos with parallel German text.
- **CSL-Daily** (Zhou et al., 2021): A Chinese Sign Language (CSL) corpus tailored for sign-to-Chinese SLT, emphasizing interactions in daily communication contexts.

Statistics of both datasets are summarizes in Table 1.

4.2 Evaluation Metrics

We evaluate our method following (Müller et al., 2022; Müller et al., 2023a), using BLEU[3] (via SacreBLEU (Post, 2018)) for lexical overlap, ROUGE (Lin, 2004)[4] for recall-oriented n-gram overlap, and BLEURT (Sellam et al., 2020)[5] for semantic quality.

4.3 State-of-the-art Systems

We evaluate our method against several strong recent state-of-the-art systems within the gloss-free paradigm. CSGCR (Zhao et al., 2021) improves SLT accuracy and fluency through three modules: word existence verification, conditional sentence generation, and cross-modal re-ranking for richer grammatical representations. GFSLT-VLP (Zhou et al., 2023) leverages vision–language pretraining, while FLa-LLM (Chen et al., 2024) adopts a two-stage gloss-free pipeline that first pre-trains the visual encoder and then fine-tunes a pre-trained LLM for SLT. Sign2GPT (Wong et al., 2024) maps visual inputs to pseudo-gloss sequences and decodes them with GPT-style language modeling, whereas SignLLM (Gong et al., 2024) discretizes sign features into visual tokens to prompt a frozen LLM. SEM-SLT (Hamidullah et al., 2024) aligns sign language videos with sentence embeddings and serves as the foundation of our work. For multilingual settings, Sign2(LID+Text) (Tan et al., 2025) combines token-level sign language identification with a CTC objective to generate spoken text.

4.4 Implementation Details

• **Feature Extraction.** We begin by processing each sign language video $x = \{f_1, f_2, \ldots, f_T\}$ as a sequence of T RGB frames. From each frame, we extract:

- **Spatial Features (s_t):** Using a Vision Transformer (ViT (Dosovitskiy et al., 2020)) pretrained on ImageNet.
- **Motion Features (m_t):** Using VideoMAE (Tong et al., 2022).

These features are then fused via the visual fusion block $\mathcal{F}$ from SpaMo to yield a joint representation h_t for each timestep.

• **Training the visual block (LoRA).** We train the visual block using LoRA with:

- **LoRA:** $r = 16,\ \alpha = 32$
- **Batching:** batch size 4, gradient accumulation 2, on 8 GPUs in parallel
- **Loss weights:**
  - $\lambda_{\text{ce}} = 0.1$ (auxiliary soft translation signal)
  - $\lambda_{\text{sem}} = 1.0$ (primary objective)
  - $\lambda_{\cos} = 2.7$ (stabilizes angular alignment)
  - $\lambda_{\text{nce}} = 0.0$
  - $\lambda_{\text{mse}} = 7000.0$ (strong magnitude regularizer)

[3] `BLEU|nrefs:1|bs:1000|seed:16|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0`

[4] `ROUGE|L|nrefs:1|tok:13a|case:mixed|version:1.5.5`

[5] BLEURT v0.0.2 using checkpoint BLEURT-20.

Because our model operates on embedding vectors with small magnitudes, the MSE loss can rapidly fall to $\sim 10^{-5}$ even when cosine similarity remains suboptimal. Empirically, we observed that **cosine and MSE only begin to correlate at** $\sim 10^{-6}$. Optimizing cosine alone often stalls, as MSE ceases to decrease, while optimizing MSE alone improves fidelity but does not guarantee angular alignment. To address this, we up-weight MSE to maintain shrinkage and retain a non-negligible cosine term to enforce directional consistency. We also experimented with InfoNCE, but under our effective batch size (with few hard negatives) it led to slower convergence and negligible improvements and we do not use it in our final experiments.

• **Sentence embedding pooling.** The original SONAR pools by running a shallow decoder: it feeds a special token (the EOS id in `M200M100`) as input and uses the encoder outputs as hidden states; the first decoder output is taken as the sentence embedding. During the Visual Block training, we adopt this approach with a shallow decoder initialized from the first three SONAR decoder layers and train it only for pooling. This supplies language context during pooling, while the incoming features themselves are language-agnostic (from another modality). Text generation is then conditioned on the target language.

• **Visual representation.** We adopt the best-performing visual representation strategies reported in prior work, noting that optimal choices vary across datasets. To ensure comparability in our multi–sign language experiments, we restrict evaluation to datasets with similar video settings and select the strongest corresponding model. The SpaMo Visual Block performs best with global, high-quality cues e.g., high-resolution videos with a moderate signer–camera distance (CSL-Daily), or lower-resolution videos where the signer is close and centered (PHOENIX-2014T). Consequently, we conduct multilingual experiments on CSL-Daily and PHOENIX-2014T.

• **Training the translation model with visual features.** We train the end-to-end translation system (with the Visual Block or the fused spatial+motion features) using the **same LoRA configuration** as above.

- **Batching & schedule:** batch size 8 on a single GPU
  - **CSL-Daily:** cosine learning-rate schedule with a peak LR of 3×10^{-4}
  - **PHOENIX-2014-T (monolingual):** constant LR (we found it more stable)
- **Text augmentation.** To expand the datasets using NLLB (NLLB Team, 2024), we machine-translate the target texts into three high-resource languages (English, French and Spanish) using the `facebook/nllb-200-distilled-600M` model.
- **Video augmentation.** Coupled with the target-language augmentation, we also perturb the input videos so that each training instance is presented with both linguistic and visual variability. At each iteration, one augmented variant is sampled. In this work we restrict ourselves to:
  - `frame_mask_ratio` $= 0.2$
  - `frame_dropout_prob` $= 0.2$
  - `add_noise_std` $= 0.04$
  - `shuffle_window` $= 3$

5 Results and Analysis

5.1 Comparative Analysis

We compare our approach with other gloss-free methods on both PHOENIX-2014T and CSL-Daily datasets in Table 2. Our method shows a clear advantage on the semantics-oriented BLEURT metric. It reaches a **BLEURT of 0.545**, outperforming the sentence-based supervision model using text-only sentence embedding (SEM-SLT). BLEURT uses a BERT-based scorer and is designed to capture meaning and fluency, unlike BLEU and ROUGE, which primarily measure n-gram overlap. Moreover, our model outperforms previous monolingual and multilingual systems on CSL-Daily in terms of BLEU and achieves comparable results on PHOENIX-2014T.

• **Observed gaps.** We observe a decrease in BLEU compared to the SEM-SLT system, which is expected since our model is not fine-tuned on sign-language text. Our language-agnostic, sentence embedding-based supervision preserves semantics without requiring fine-tuning on specific dataset: it goes beyond surface n-gram matching to produce translations that are contextually accurate, grammatically correct, and cross-lingually robust. Part

Method	PHOENIX-2014T			CSL-Daily		
	BLEU	BLEURT	RG	BLEU	BLEURT	RG
Monolingual						
CSGCR (Zhao et al., 2021)	15.18	–	38.85	–	–	–
GFSLT-VLP (Zhou et al., 2023)	21.44	–	42.29	11.00	–	36.44
FLa-LLM (Chen et al., 2024)	23.09	–	45.27	14.20	–	37.25
Sign2GPT (Wong et al., 2024)	22.52	–	48.90	15.40	–	**42.36**
SignLLM (Gong et al., 2024)	23.40	–	44.49	15.75	–	39.91
SEM-SLT (Hamidullah et al., 2024)	24.10	0.481	–	–	–	–
Multilingual						
Sign2(LID+Text) (Tan et al., 2025)	**24.23**	–	**50.60**	14.18	–	40.00
SONAR-SLT (Ours)	22.01	**0.545**	41.44	16.23	**0.561**	42.29

Table 2: Comparison of SONAR-SLT with other gloss-free models on PHOENIX-2014T and CSL-Daily (metrics: BLEU, BLEURT, ROUGE (RG)). Unreported metrics are left blank; SONAR-SLT sets the best reported BLEURT on PHOENIX-2014T and remains strongly competitive with several LLM-based baselines on both datasets.

Resource	Language	BLEU
High	Spanish (es)	22.3
	French (fr)	22.6
	English (en)	21.6
Low	Turkish (tr)	13.1
	Malagasy (mg)	11.8
	Persian (fa)	8.7

Table 3: SONAR-SLT performance across target languages in both high- and low-resource settings on PHOENIX-2014T, reported using BLEU scores.

of the remaining gap stems from dataset capture conditions. Our feature extractor (Hwang et al., 2025) is tuned for global cues and can be less accurate in cases where fine-grained articulations, such as facial expressions and finger movements, are critical. Recent top systems address this with keypoint-based representations and extensive preprocessing (Gueuwou et al., 2025b), which help preserve these fine-grained details.

• **Multilingual and multi–sign language.** We evaluate target-side augmentation, where language translations are included in training. Results for both low- and high-resource languages on PHOENIX-2014T are presented in Table 3. In our experiments, we augmented the target set in training with three high-resource languages—French, Spanish, and English—while the model was evaluated on other unseen languages. Using this augmented target set yields a modest improvement over training with a single target language. However, we observe a gap in performance between high- and low-resource languages, which primarily stems from lower reference translation quality in the low-resource languages.[6] The narrow domain of PHOENIX-2014T can also introduce dataset-specific idiosyncrasies, complicating fair comparisons.

Table 4 shows that pre-training on concatenated multi-sign corpora followed by monolingual fine-tuning proves most effective. In contrast, joint multi-sign fine-tuning risks resembling another full training run without yielding substantial gains. In our experiments, we first pre-train on the combined data and then fine-tune monolingually, consistent with (Hamidullah et al., 2024); post-fine-tuning performance remains largely unchanged (see Table 4, mono vs. multilingual setup). Differences in dataset capture conditions still matter—for example, methods that rely solely on global visual features can underperform when fine-grained articulations, such as hand or facial details, are crucial. Pipelines that integrate keypoints with extensive preprocessing (Gueuwou et al., 2025b) help mitigate such losses and achieve stronger results.

5.2 Multitask Learning Effect on the Visual Block

The effect of sentence-embedding supervision is strongest when the Visual Block is still learning feature representations. Once the block has converged—or is pretrained—the additional impact of cosine or MSE objectives diminishes. This occurs because cross-entropy loss often remains rel-

[6] Reference translations were obtained using `facebook/nllb-200-distilled-600M` model.

Type	Variant	PHOENIX-2014T			CSL-Daily		
		BLEURT	BLEU	RG	BLEURT	BLEU	RG
Multi	VB pretrained	**0.523**	21.52	41.10	**0.561**	**16.23**	**42.29**
	VB scratch	0.508	21.38	**42.03**	0.472	14.68	42.12
	VB frozen	0.516	**21.56**	41.39	0.549	16.06	41.95
Mono	VB pretrained	**0.545**	**22.01**	40.52	**0.558**	**16.07**	**42.13**
	VB scratch	0.490	19.79	39.95	0.447	14.14	40.59
	VB frozen	0.520	21.56	**41.44**	0.529	15.70	41.79

Table 4: SONAR-SLT results for the Visual Block (VB) variants under Multilingual and Monolingual settings on PHOENIX-2014T and CSL-Daily. Metrics include BLEURT, BLEU, and ROUGE (RG); best scores per dataset/metric are in bold.

atively high (above 2–3), while MSE rapidly falls to $\mathcal{O}(10^{-5})$ and cosine similarity saturates around ~ 0.3.

In contrast, introducing the auto-encoding loss provides a second cross-entropy signal, which exerts a stronger influence on the Visual Block. Here, intermediate supervision continues to be beneficial, and the auto-encoding objective itself accelerates convergence. We consistently observed this effect in CSL-Daily and in the augmented translation setup on PHOENIX-2014T.

5.3 Qualitative and Semantic Error Analysis

• **Qualitative analysis.** Table 5 shows examples of two contrasting outcomes: cases where the model accurately captures the intended meaning and cases where it fails. When contextual understanding is incomplete, the decoder frequently compensates by generating fluent continuations via next-token prediction. This behavior is characteristic of SLT systems that rely on pretrained language models as decoders: they can mask weaknesses in semantic grounding by producing outputs that are coherent but only partially faithful to the source. As a result, improvements in BLEU may reflect the decoder's ability to recover plausible sentences rather than true gains in sign-to-text comprehension. Therefore, exact sequence matching metrics such as BLEU are insufficient and in some cases misleading for evaluating translation quality in SLT.

Language-specific tendencies. We deep into the analysis of two languages: German, a language trained with original data and French, a language trained via machine translation augmentation.

- **German:** Errors often arise from compound nouns, flexible word order, and embedded clauses, leading to partial omissions, attribute reordering, or unnatural compounds. When alignment is uncertain, the model may insert generic stock phrases or repetitions.

- **French:** Our analysis shows more frequent noun substitutions, agreement mismatches, and text modality shifts (e.g., hedging with *"sont possibles"*). Register differences from determiners or prepositions are also common. Incorrect date and numeric substitutions occur more frequently than in German, likely due to segmentation differences in temporal expressions.

• **Semantic analysis.** Surface-form scores vs. meaning preservation. We observe a systematic mismatch between surface-form metrics (e.g., BLEU) and semantic adequacy (BLEURT) across both German and French. Outputs with only moderate n-gram overlap can still be semantically faithful, while some high-scoring predictions contain factual errors.

Semantically near correct and correct paraphrases (German). As illustrated by the green-highlighted examples in Table 5, incorrect lexical or numeric substitutions leave most of the remaining meaning intact (e.g., date shifts: *"Sonntag, den neunzehnten Dezember"* → *"Sonntag, den siebzehnten August"*; temperature adjustments: *"sechs Grad an den Alpen"* → *"neun Grad am Alpenrand"*). We also observe benign stylistic reformulations (*"es gelten entsprechende Warnungen"* → *"es bestehen Unwetterwarnungen"*) and word-order changes without semantic effect (*"aus Südwest bis West"* → *"aus West bis Südost"*).

Semantically near correct and correct paraphrases (French). Similarly, the first green-highlighted examples show structural or modality shifts that preserve much of the remaining meaning,

German (DGS weather domain)
Ref (DE): Sonntag, den neunzehnten Dezember. *EN: Sunday, the nineteenth of December.* **Pred (DE):** Sonntag, den siebzehnten August. *EN: Sunday, the seventeenth of August.*
Ref (DE): sechs Grad an den Alpen. *EN: Six degrees in the Alps.* **Pred (DE):** neun Grad am Alpenrand. *EN: Nine degrees on the edge of the Alps.*
Ref (DE): Höhenlagen Süddeutschlands. *EN: High-altitude areas of southern Germany.* **Pred (DE):** Küsten. *EN: Coasts.*
French (weather domain)
Ref (FR): vingt-huit août. *EN: Twenty-eighth of August.* **Pred (FR):** vingt-cinq novembre. *EN: Twenty-fifth of November.*
Ref (FR): Des rafales orageuses de l'ouest. *EN: Stormy gusts from the west.* **Pred (FR):** Des rafales orageuses sont possibles. *EN: Stormy gusts are possible.*
Ref (FR): Risque d'inondation. *EN: Risk of flooding.* **Pred (FR):** Avertissements météorologiques violents. *EN: Severe weather warnings.*

Table 5: German and French examples —two semantically near correct paraphrases (green) and one semantically incorrect output (red), with English translations.

such as date substitutions (*"vingt-huit août"* → *"vingt-cinq novembre"*), modality changes (*"Des rafales orageuses de l'ouest"* → *"Des rafales orageuses sont possibles"*), or expanded phrasing (*"Également orages sur la mer du Nord"* → *"Il y a également des orages sur la mer du Nord"*).

Semantically incorrect outputs, true errors (French and German). The red-highlighted rows in Table 5 illustrate errors such as topic drift (predicting wind instead of temperature), incorrect locations (*"Höhenlagen Süddeutschlands"* → *"Küsten"*; *"sud-est"* → *"nord"*), system inversions (*"Hoch"* ↔ *"Tief"*; *"haut"* ↔ *"profonde"*), hallucinated entities, or incorrect hazard categories (*"Risque d'inondation"* → *"Avertissements météorologiques violents"*).

Overall, as in other machine translation tasks, n-gram metrics penalize near or even fully legitimate paraphrases and sometimes fail to capture serious factual errors. Robust SLT evaluation requires *semantic* metrics that explicitly reward meaning preservation while penalizing distortions or hallucinations.

Implications. Evaluation and model development for multilingual SLT should be language-aware. In practice, one should combine semantics-focused metrics with targeted, language-specific checks (e.g., temporals and agreement in French; word order and compounding in German) to obtain fair comparisons and actionable diagnostics.

6 Conclusion

We presented a scalable SLT framework that breaks the traditional close dependency between sign and spoken languages in training data and system development. By aligning sign language videos with multilingual, multimodal sentence embeddings from SONAR, our approach yields a language-agnostic semantic representation that generalizes across both sign languages and spoken targets. This reduces reliance on language-model priors and prioritizes visual grounding and SLT-specific grammar over surface-level text patterns.

Experiments show that language-agnostic supervision enables robust translation even under sign–target mismatches. Multilingual text augmentations, combined with visual augmentation, improves performance on PHOENIX-2014T despite limited data. Ablations further confirm the advantages of this approach in preserving semantic adequacy.

Current evaluation practices often emphasize surface overlap rather than meaning. Future work should develop metrics aligned with semantic similarity and extend supervision to low-resource sign languages and continuous signing in the wild.

Limitations

Our main limitation lies in the visual feature extractor rather than the model architecture itself. We used a pre-existing visual block to avoid evaluation bias, which restricted us to datasets with compatible video settings (CSL-Daily and PHOENIX-2014T) and excluded larger corpora such as How2Sign or YouTube-ASL. As a result, our approach focuses on preserving semantics rather than maximizing exact sentence matches.

Machine translation for data augmentation might induce unintended cultural mistakes that go beyond literal translation. The evaluation on non-human translated datasets also limits the strength of the conclusions for low-resourced languages.

Acknowledgments

This work was partially funded by the German ministry for education and research (BMBF) through projects BIGEKO (grant number 16SV9093) and TRAILS (grant number 01IW24005).

References

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural sign language translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Zhigang Chen, Benjia Zhou, Jun Li, Jun Wan, Zhen Lei, Ning Jiang, Quan Lu, and Guoqing Zhao. 2024. Factorized learning assisted with large language model for gloss-free sign language translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7071–7081, Torino, Italia. ELRA and ICCL.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *ArXiv*.

Jia Gong, Lin Geng Foo, Yixuan He, Hossein Rahmani, and Jun Liu. 2024. Llms are good sign language translators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18362–18372.

Mo Guan, Yan Wang, Guangkun Ma, Jiarui Liu, and Mingzu Sun. 2025. Mska: Multi-stream keypoint attention network for sign language recognition and translation. *Pattern Recogn.*, 165(C).

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, and Karen Livescu. 2025a. SignMusketeers: An efficient multi-stream approach for sign language translation at scale. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22506–22521, Vienna, Austria. Association for Computational Linguistics.

Shester Gueuwou, Xiaodan Du, Greg Shakhnarovich, Karen Livescu, and Alexander H. Liu. 2025b. SHuBERT: Self-supervised sign language representation learning via multi-stream cluster prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28792–28810, Vienna, Austria. Association for Computational Linguistics.

Shester Gueuwou, Sophie Siake, Colin Leong, and Mathias Müller. 2023. JWSign: A highly multilingual corpus of Bible translations for more diversity in sign language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9907–9927, Singapore. Association for Computational Linguistics.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2022. Spatio-temporal sign language representation and translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 977–982.

Yasser Hamidullah, Josef van Genabith, and Cristina España-Bonet. 2024. Sign language translation with sentence embedding supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 425–434, Bangkok, Thailand. Association for Computational Linguistics.

Eui Jun Hwang, Sukmin Cho, Junmyeong Lee, and Jong C. Park. 2025. An efficient gloss-free sign language translation using spatial configurations and motion dynamics with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3901–3920, Albuquerque, New Mexico. Association for Computational Linguistics.

Zecheng Li, Wengang Zhou, Weichao Zhao, Kepeng Wu, Hezhen Hu, and Houqiang Li. 2025. Uni-sign: Toward unified sign language understanding at scale. *Preprint*, arXiv:2501.15187.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation*, pages 68–94, Singapore. Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden,

Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2022. Findings of the First WMT Shared Task on Sign Language Translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation*, pages 744–772, Abu Dhabi. Association for Computational Linguistics.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sihan Tan, Taro Miyazaki, and Kazuhiro Nakadai. 2025. Multilingual gloss-free sign language translation: Towards building a sign language foundation model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 553–561, Vienna, Austria. Association for Computational Linguistics.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-ASL: A large-scale, open-domain american sign language-english parallel corpus. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Ryan Wong, Necati Cihan Camgoz, and Richard Bowden. 2024. Sign2GPT: Leveraging large language models for gloss-free sign language translation. In *The Twelfth International Conference on Learning Representations*.

Shakib Yazdani, Josef Van Genabith, and Cristina España-Bonet. 2025a. Continual learning in multilingual sign language translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10923–10938, Albuquerque, New Mexico. Association for Computational Linguistics.

Shakib Yazdani, Yasser Hamidullah, Cristina España-Bonet, and Josef van Genabith. 2025b. Seeing, signing, and saying: A vision-language model-assisted pipeline for sign language data acquisition and curation from social media. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI era*, pages 1374–1384, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aoxiong Yin, Zhou Zhao, Weike Jin, Meng Zhang, Xingshan Zeng, and Xiaofei He. 2022. Mlslt: Towards multilingual sign language translation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5109.

Jian Zhao, Weizhen Qi, Wengang Zhou, Nan Duan, Ming Zhou, and Houqiang Li. 2021. Conditional sentence generation and cross-modal reranking for sign language translation. *IEEE Transactions on Multimedia*, 24:2662–2672.

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, and Du Zhang. 2023. Gloss-free sign language translation: Improving from visual-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20871–20881.

Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. 2021. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1325.

A Appendix

A.1 Detailed Architecture

Figure 3 shows the details of our system architecture as explained in Section 3.

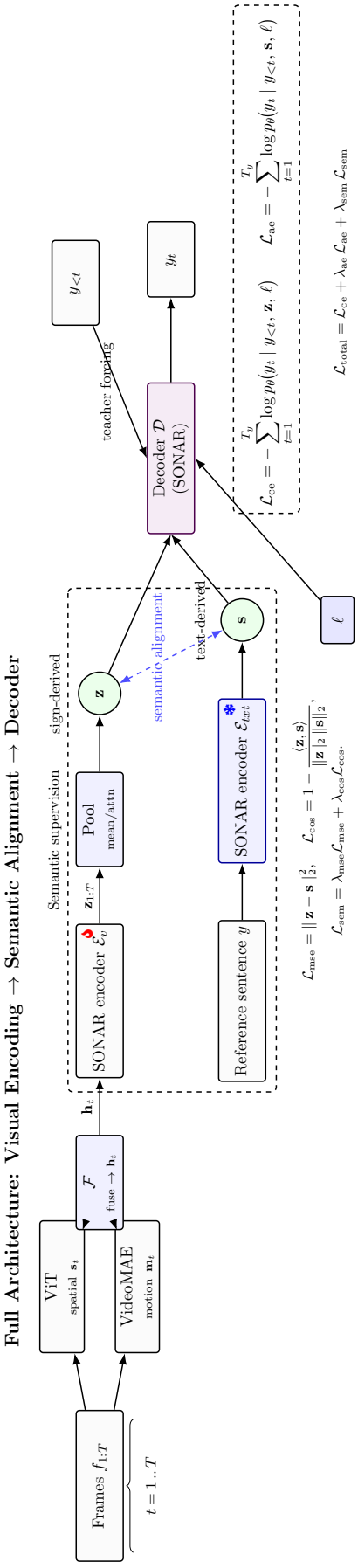


Figure 3: Detailed architecture without the contrastive term (NCE loss).

A.2 Porting SONAR from NLLB Fairseq to Huggingface.

SONAR is officially supported in fairseq, but only its text encoder is available on Hugging Face. To enable full conditional generation, we ported both the encoder and decoder weights from the original SONAR checkpoints into `M200M100`, extending the earlier encoder-only port provided by the NLLB team. In particular, we transferred the decoder weights directly from fairseq, validated their functionality, and released the complete model for public use. The resulting `M200M100ForConditionalGeneration` can now be loaded end-to-end and fine-tuned directly.

A.3 Additional Qualitative Results

Additional translation examples for CSL-Daily and PHOENIX-2014T are provided in Tables 6 and 7, respectively.

Text
Ref (ZH): 我们下午三点见面。 *EN: We will meet at three in the afternoon.* **Pred (ZH):** 我们三点钟下午见。 *EN: We meet at three o'clock in the afternoon.*
Ref (ZH): 我早上吃面包和牛奶。 *EN: I eat bread and milk in the morning.* **Pred (ZH):** 我早上吃了牛奶和面包。 *EN: I had milk and bread in the morning.*
Ref (ZH): 我们乘坐飞机去旅游，今天在酒店住宿。 *EN: We took a plane to travel, and are staying in a hotel today.* **Pred (ZH):** 我们飞机去上海，今天喝酒睡觉。 *EN: We took a plane to Shanghai, today we drink alcohol and sleep.*

Table 6: CSL-Daily examples —good translations (green) and one bad translation (red), showing reference and prediction in Chinese, with English translations for clarity.

Text
Ref (DE): ich wünsche ihnen noch einen schönen abend. *EN: I wish you a pleasant evening.* **Pred (DE):** und jetzt wünsche ich ihnen noch einen schönen abend. *EN: And now, I wish you a pleasant evening.* **Pred (FR):** Et maintenant, je vous souhaite une bonne soirée. *EN: And now, I wish you a good evening.*
Ref (DE): der wind aus süd bis west weht schwach bis mäßig. *EN: The wind from the south to west blows weakly to moderately.* **Pred (DE):** der wind weht meist schwach aus süd bis west. *EN: The wind generally blows weakly from south to west.* **Pred (FR):** Le vent souffle généralement faiblement du sud-ouest. *EN: The wind generally blows weakly from the southwest.*
Ref (DE): in deutschland gibt es nur schwache luftdruckunterschiede. *EN: In Germany, there are only slight air pressure differences.* **Pred (DE):** im nordosten deutschlands sorgt das hoch für wenig unbeständiges wetter. *EN: In northeastern Germany, the high pressure system causes little unsettled weather.* **Pred (FR):** Dans certaines régions de l'Allemagne, la pression atmosphérique élevée n'est toujours pas atteinte. *EN: In some regions of Germany, high atmospheric pressure has still not been reached.*

Table 7: PHOENIX-2014T examples —two good translations (green) and one bad translation (red), showing reference (German), predictions (German and French), and English translations.

GAMBIT+: A Challenge Set for Evaluating Gender Bias in Machine Translation Quality Estimation Metrics

Giorgos Filandrianos[1,2*] **Orfeas Menis Mastromichalakis**[1] **Wafaa Mohammed**[3]
Giuseppe Attanasio[2] **Chrysoula Zerva**[2,4,5]
[1]National Technical University of Athens, Greece
[2]Instituto de Telecomunicações, Lisbon, Portugal [3]University of Amsterdam, Netherlands
[4]Instituto Superior Técnico, Universidade de Lisboa, Portugal [5]ELLIS Unit Lisbon, Portugal

Abstract

Gender bias in machine translation (MT) systems has been extensively documented, but bias in automatic quality estimation (QE) metrics remains comparatively underexplored. Existing studies suggest that QE metrics can also exhibit gender bias, yet most analyses are limited by small datasets, narrow occupational coverage, and restricted language variety. To address this gap, we introduce a large-scale challenge set specifically designed to probe the behavior of QE metrics when evaluating translations containing gender-ambiguous occupational terms. Building on the GAMBIT corpus of English texts with gender-ambiguous occupations, we extend coverage to three source languages that are genderless or natural-gendered, and eleven target languages with grammatical gender, resulting in 33 source–target language pairs. Each source text is paired with two target versions differing only in the grammatical gender of the occupational term(s) (masculine vs. feminine), with all dependent grammatical elements adjusted accordingly. An unbiased QE metric should assign equal or near-equal scores to both versions. The dataset's scale, breadth, and fully parallel design, where the same set of texts is aligned across all languages, enables fine-grained bias analysis by occupation and systematic comparisons across languages.

1 Introduction

While gender bias in machine translation systems is widely acknowledged and has been widely documented, the biases in the translations of gender-ambiguous terms, such as occupational titles, is a topic relatively unexplored, which has lately gained some traction (Mastromichalakis et al., 2025). These biases are often reflected in the disproportionate assignment of a gender (e.g. masculine forms) to occupations when the gender of the subject is unknown. While overall much research has focused on bias in MT outputs, comparatively little attention has been paid to the potential gender biases of automatic quality estimation metrics. Recent studies suggest that QE metrics, which are intended to provide an objective measure of translation quality, can also exhibit systematic biases (Zaranis et al., 2024). However, existing analyses on gender-ambiguous inputs are typically limited by datasets with short texts, typically sentence-level, a restricted variety of occupations, and a limited range of language pairs.

To address these limitations, we introduce GAMBIT+[1], a large-scale challenge set specifically designed to evaluate gender bias in QE metrics when translating gender-ambiguous occupational terms. Our dataset extends the original GAMBIT corpus[2] (Mastromichalakis et al., 2025) of English texts with gender-ambiguous occupations to include three source languages—English (a natural-gendered language), Turkish and Finnish (both genderless languages)—and eleven target languages with grammatical gender: Arabic, Czech, Greek, Spanish, French, Icelandic, Italian, Portuguese, Russian, Serbian, and Ukrainian. This results in 33 source–target language pairs, covering a wide spectrum of linguistic typologies and providing a rich resource for cross-linguistic analysis of gender bias in QE.

For each source text, we provide two target versions differing only in the grammatical gender of the occupational term (masculine vs. feminine). All necessary adjustments for grammatical correctness, such as gendered adjectives, are applied consistently across both versions. An unbiased QE metric should assign equal or near-equal scores to the two versions, as there is no indication of gender in the source text. Unlike prior work, which often aggregates bias analysis across all occupations or texts in general, GAMBIT+ explicitly tracks the occupational terms mentioned in each text and links them to ISCO-08 codes (hereafter, ISCO codes) [3]. This enables fine-grained, occupation-level

*Corresponding author: geofila@ails.ece.ntua.gr.

[1]GAMBIT+ is available at https://huggingface.co/datasets/ailsntua/gambit-plus.

[2]https://huggingface.co/datasets/ailsntua/GAMBIT

[3]ISCO-08 is an internationally recognized system for

Proceedings of the Tenth Conference on Machine Translation, pages 314–326
November 8-9, 2025

analyses and facilitates the study of stereotypical patterns, as previous studies have shown that certain occupations are more likely to be translated in a gendered manner according to societal stereotypes (Mastromichalakis et al., 2025; Menis-Mastromichalakis et al., 2025).

The parallel design of the dataset ensures that all source texts are aligned across target languages and both gendered versions, enabling consistent and controlled benchmarking of QE metrics. With this resource, researchers can investigate not only overall tendencies of QE systems but also nuanced, occupation-specific, and cross-lingual patterns, providing a more comprehensive understanding of how gender bias manifests in translation evaluation.

As part of the challenge set subtask of the shared task on Automated Translation Quality Evaluation Systems at WMT 2025 (Lavie et al., 2025), GAMBIT+ was used to evaluate a set of established QE metrics as well as participant submissions to the shared task. This allowed us to benchmark the performance of different metrics in a controlled, fine-grained setting and to assess how gender bias manifests in real-world QE systems. In this paper, we present and discuss the results of these evaluations, highlighting both systematic tendencies and occupation-specific patterns, and providing insights into the current limitations and strengths of existing QE approaches with respect to gender fairness.

2 Related Work

Gender bias in Machine Translation has been widely documented, revealing persistent disparities influenced by societal norms, model design choices, and deployment contexts (Savoldi et al., 2021; Vanmassenhove, 2024; Savoldi et al., 2024; Menis-Mastromichalakis et al., 2025; Mastromichalakis et al., 2025). Numerous studies have examined the prevalence and consequences of such biases across languages, cultural settings, and MT architectures (Rescigno et al., 2020; Paolucci et al., 2023; Farkas and Németh, 2022; Ghosh and Caliskan, 2023; Kostikova et al., 2023; Piazzolla et al., 2023), underscoring the need for robust evaluation and mitigation. A particularly relevant line of work focuses on occupational bias in MT, where stereotypes associated with specific professions affect translation choices (Gorti et al., 2024; Tal et al., 2022; Mastromichalakis et al., 2025). Related efforts in other NLP tasks have explored gender ambiguity in Question Answering (Parrish et al., 2022; Li et al., 2020) and coreference resolution (Rudinger et al., 2018; Zhao et al., 2018; Kotek et al., 2023). In MT, proposed strategies for handling ambiguity include generating all grammatically correct gendered translations (Garg et al., 2024), and disambiguating inputs before translation (Vanmassenhove et al., 2018). Resources supporting these investigations range from knowledge graphs (Mastromichalakis et al., 2024) to multilingual benchmarks and challenge sets (Currey et al., 2022). Mitigation strategies have included model fine-tuning, data balancing, and adaptive learning (Saunders and Byrne, 2020; Escudé Font and Costa-jussà, 2019; Costa-jussà and de Jorge, 2020), as well as gender-neutral translation approaches (Piergentili et al., 2023a; Lardelli and Gromann, 2023) and benchmarks for evaluating them (Piergentili et al., 2023b; Lardelli et al., 2024; Gkovedarou et al., 2025). A central element in all these efforts is evaluation, since quality estimation metrics determine what counts as a "good" translation and thus influence MT system development.

While several studies have examined whether evaluation metrics for natural language generation exhibit social biases—such as Qiu et al. (2023), who compared n-gram- and model-based metrics, Sun et al. (2022), who quantified different bias types, and Gao and Wan (2022), who measured race and gender stereotypes—very few works have explored gender bias in MT QE metrics. One exception is Zaranis et al. (2024), which conducted a multifaceted analysis of QE methods and demonstrated that the metrics themselves can be biased, potentially perpetuating the very stereotypes they are used to assess.

A key limitation of existing work on gender bias in QE is the lack of suitable datasets, particularly for studying gender ambiguity. Many challenge sets, such as WinoMT (Stanovsky et al., 2019) and MT-GenEval (Currey et al., 2022), focus on cases where gender can be resolved via coreference or other contextual cues. The MuST-SHE corpus (Bentivogli et al., 2020) similarly contains audio and textual cues that reveal gender (e.g., speaker's voice, pronouns, named entities). Conversely, datasets containing true gender ambiguity (Rudinger et al., 2018; Zhao et al., 2018) are often limited in language coverage, consist of isolated sentences, and lack occupational diversity, restricting fine-grained analysis. In contrast, our challenge set builds on GAMBIT (Mastromichalakis et al., 2025), where occupational mentions are inherently gender-ambiguous and no gold-standard gendered translation

classifying occupations endorsed by the International Labour Organisation (ILO). It provides a hierarchical structure that categorizes jobs into four levels of increasing granularity, using a digit-based coding system. More details: https://ilostat.ilo.org/methods/concepts-and-definitions/classification-occupation/

exists. This enables us to test whether QE metrics assign different scores to two theoretically equivalent translations differing only in the grammatical gender of the occupation, offering a controlled setting for investigating metric bias in ambiguous contexts.

3 Challenge Set Creation

The creation of the challenge set, GAMBIT+, was grounded in the GAMBIT dataset (Mastromichalakis et al., 2025), which covers all ISCO occupations (all 4-digit codes) and contains the corresponding ISCO codes and names, with English texts distributed evenly across five formats: short stories, brief news reports, short statements, short conversations, and short presentations, in which the given occupation appears. In these texts, the gender of the occupation is not explicitly indicated. For example, in the sentence "The professor delivered an engaging lecture on generation modification by the auditorium", shown in Table 1, the occupation "professor" is mentioned without any linguistic cues that would reveal its gender.

In GAMBIT+, these gender-neutral English sentences were translated into gendered target languages, producing parallel masculine and feminine versions of each sentence while preserving all other aspects of the translation. In addition, we extended the source language coverage to include genderless languages such as Turkish and Finnish, alongside English, ensuring a diverse and balanced dataset. In total, the challenge set comprises 29,415 source instances (9,805 each of the three source languages), which were translated into 11 target languages, resulting in 323,565 triplets, each consisting of the source sentence with a masculine and a feminine translation. Table 1 presents an example from the GAMBIT+ dataset with the data from GAMBIT and the corresponding entries created for the Challenge Set. Specifically, the following sections describe the generation process for producing these translations and the evaluation procedure applied to assess and ensure the high quality of the resulting dataset.

3.1 Generation

Gendered Languages. The construction of the challenge set for gendered languages was based on the English texts from the GAMBIT dataset in which the gender of the occupation was ambiguous. These sentences served as the source material for the initial stage of the process. An LLM[4] was subsequently instructed to translate each sentence into the target language, specifically translating the occupation into a given gender form (e.g., masculine). Following this, a separate interaction with the same model was initiated, ensuring that no conversational history from the first stage was preserved, and the model was instructed to take the translated text and produce a variation in the target language that preserved the text exactly, modifying only the gender of the occupation from the initial to the alternate form (e.g., feminine respectively).

This process yielded three aligned versions for each source sentence:

1. the original English text s containing a gender-ambiguous occupation,
2. the translation with the occupation in its masculine form t_{male}, and
3. the translation with the occupation in its feminine form t_{female}.

The masculine and feminine translations were constructed to be semantically and lexically identical, differing only in the gender marking of the occupation and the dependent grammatical elements. There were a few cases where the masculine and feminine variations were identical in some target languages, mostly due to the morphology of the source text and the grammatical restrictions of the target languages. All such examples were discarded from all languages (so if the masculine and feminine translations were identical in at least one language, this text was removed from all languages) in order to keep the challenge set consistent, and the texts in all languages aligned. After filtering, we retained 8,771 samples per language pair, resulting in a total of 289,443 samples in the GAMBIT+ dataset.

Genderless Languages. For genderless languages, a similar approach was applied. In this case, the model was instructed to translate the original text into the genderless target language without introducing any linguistic cues that might imply a specific gender for the occupation. This step was designed to avoid subtle cases where unintended gender hints could arise, even in languages without grammatical gender. Since in those languages there is no grammatical gender, the second step that generated the gendered variation in gendered languages was not applied.

By following this procedure, we obtained aligned texts across multiple languages, comprising source languages with gender-ambiguous occupations and target languages with matched masculine and feminine translations for the same occupations, as shown

[4]The LLM used is Claude Sonnet 3.5 v2

GAMBIT			
ISCO ID	**ISCO Name**	**Type**	**English Text**
2310	Professor	short statement	The Professor delivered an engaging lecture on quantum mechanics to a packed auditorium.

GAMBIT+ Additional Source Languages	
Language	**Text**
Turkish	Profesör, tıklım tıklım dolu bir konferans salonunda kuantum mekaniği üzerine etkileyici bir ders verdi.
Finnish	Professori piti mukaansatempaavan luennon kvanttimekaniikasta täydelle luentosalille.

GAMBIT+ Target Languages		
Language	**Masculine**	**Feminine**
Arabic	ألقى الأستاذ محاضرة مشوقة عن الميكانيكا الكمية في قاعة محاضرات مكتظة.	ألقت الأستاذة محاضرة مشوقة عن الميكانيكا الكمية في قاعة محاضرات مكتظة.
Greek	Ο Καθηγητής έδωσε μια συναρπαστική διάλεξη για την κβαντική μηχανική σε ένα κατάμεστο αμφιθέατρο.	Η Καθηγήτρια έδωσε μια συναρπαστική διάλεξη για την κβαντική μηχανική σε ένα κατάμεστο αμφιθέατρο.
Czech	Profesor přednesl poutavou přednášku o kvantové mechanice před zaplněným auditoriem.	Profesorka přednesla poutavou přednášku o kvantové mechanice před zaplněným auditoriem.
Icelandic	Herra prófessorinn flutti áhugaverðan fyrirlestur um skammtafræði fyrir fullum fyrirlestrasal.	Frú prófessorinn flutti áhugaverðan fyrirlestur um skammtafræði fyrir fullum fyrirlestrasal.
Italian	Il Professore ha tenuto una coinvolgente lezione sulla meccanica quantistica in un auditorium gremito.	La Professoressa ha tenuto una coinvolgente lezione sulla meccanica quantistica in un auditorium gremito.
Russian	Профессор прочитал увлекательную лекцию по квантовой механике в переполненной аудитории.	Профессор прочитала увлекательную лекцию по квантовой механике в переполненной аудитории.
French	Le Professeur a donné une conférence passionnante sur la mécanique quantique dans un amphithéâtre comble.	La Professeure a donné une conférence passionnante sur la mécanique quantique dans un amphithéâtre comble.
Spanish	El Profesor dio una conferencia cautivadora sobre mecánica cuántica ante un auditorio repleto.	La Profesora dio una conferencia cautivadora sobre mecánica cuántica ante un auditorio repleto.
Portuguese	O Professor ministrou uma palestra envolvente sobre mecânica quântica para um auditório lotado.	A Professora ministrou uma palestra envolvente sobre mecânica quântica para um auditório lotado.
Serbian	Profesor je održao zanimljivo predavanje o kvantnoj mehanici pred punim auditorijumom.	Profesorica je održala zanimljivo predavanje o kvantnoj mehanici pred punim auditorijumom.
Ukrainian	Професор провів захопливу лекцію з квантової механіки в переповненій аудиторії.	Професорка провела захопливу лекцію з квантової механіки в переповненій аудиторії.

Table 1: Example instance from GAMBIT+ dataset, showing GAMBIT metadata, source translations, and target translations (masculine/feminine).

in the example of Table 1. The model used for the generation is Claude Sonnet 3.5 v2 [5].

3.2 Evaluation

Ensuring that the translations differed solely in the gender marking of the occupation, while remaining otherwise semantically equivalent, was a central requirement of the data creation process. However, adherence to this constraint could not be assumed, even though the model had been explicitly prompted to preserve all other aspects of the source text. To verify compliance, an additional evaluation was conducted in an LLM-as-a-judge approach (Gu et al., 2024). A different and more capable model[6], namely Claude Sonnet 3.7 [7], was provided with two texts: the translation in the masculine form and the translation in the feminine form. The model was instructed to analyse the two sentences and determine whether the only difference between them was the gender of the occupation, or whether additional semantic or structural differences were present. This evaluation was carried out for all language-pair combinations using 10% of the dataset due to computation constraints, and the results are reported in Table 2.

As we can see, in almost all cases the accuracy is above 90%, indicating a high quality of generated data. Additionally, we manually reviewed around 100 samples of the "errors" identified by the LLM judge to further investigate the issues. In most cases, the errors flagged by the judge were not truly errors but rather stemmed from minor variations in dependent gram-

[5] anthropic.claude-3-5-sonnet-20241022-v2:0

[6] https://www.anthropic.com/news/claude-3-7-sonnet

[7] anthropic.claude-3-7-sonnet-20250219-v1:0

Lang.	Acc. %	Lang.	Acc. %
Portuguese	98.38	Russian	94.00
Spanish	97.62	Ukrainian	92.62
French	97.25	Czech	90.88
Italian	96.50	Serbian	90.62
Arabic	96.25	Icelandic	86.75
Greek	95.62		

Table 2: Gender-only differences between masculine and feminine translations across languages, as assessed by an LLM-as-a-judge approach.

matical elements, which should indeed be adjusted to match the gender changes of the occupation. Such differences often led the judge to conclude that the outputs were not aligned, revealing a certain oversensitivity in the evaluation. As a result, when the metric marked something as incorrect, it was not necessarily a substantive error, whereas we observed that when it marked something as correct, this judgment was indeed accurate. Consequently, the reported numbers are likely stricter than the actual performance.

The prompts used for data generation and evaluation are provided in Appendix A.

4 Evaluation Setup & Analysis Approach

For the evaluation, we prepared a dataset in which each entry consisted of a source text (s) in one of three source languages (English, Turkish, Finnish) and, for each target language, two corresponding hypotheses: one where the occupation appeared in the translated masculine form (t_{male}) and another where it appeared in the translated feminine form (t_{female}). The two hypotheses were constructed to differ solely in the grammatical gender marking of the occupation, with all other aspects of the sentence kept identical. This dataset was then provided by the shared task organizers to the participating teams, who ran their metrics and returned their scores for each hypothesis in every source–target language pair, while the organizers themselves used a set of baseline metrics on our dataset too. These returned results form the basis of the subsequent analysis. The source languages were 3 (English, Turkish, Finnish) and the target languages were 11 (Arabic, Czech, Greek, Spanish, French, Icelandic, Italian, Portuguese, Russian, Serbian, Ukrainian), leading to a total of 33 distinct source–target language pairs.

The organizers returned results for 11 evaluation metrics, 3 baseline metrics (sentinel-cand (Perrella et al., 2024), sentinel-src (Perrella et al., 2024), and COMETKiwi22 (Rei et al., 2022)), and 8 submissions (UvA-MT (Wu and Monz, 2025), rankedCOMET (Maharjan and Shrestha, 2025), MetricX-25-QE (Juraska et al., 2025), MetricX-25 (Juraska et al., 2025), baseCOMET (Maharjan and Shrestha, 2025), Polycand-1 (Züfle et al., 2025), Polycand-2 (Züfle et al., 2025), and Polyic-3 (Züfle et al., 2025)). All metrics, except for UvA-MT, were run on the complete set of 33 source-target language pairs. The UvA-MT metric was evaluated only with English as the source language, covering all corresponding target languages. sentinel-src relies exclusively on the input, which results in identical scores being assigned to both gendered forms; therefore, it is excluded from our subsequent analyses

5 Results and Discussion

5.1 Experimental Setup

Let $M(s,t) \in [\min(M), \max(M)]$ be a translation quality metric, where s denotes the source text, t the translated text, and $M(s,t)$ returns a real-valued score within the metric's range. Our goal is to quantify the variation in M when the translations differ *only* in the grammatical gender of an occupation, while all other elements of the text remain identical.

For each source sentence s, we consider two translations: t_{male} (masculine form) and t_{female} (feminine form). The *absolute difference* in metric scores is defined as:

$$\Delta_{\text{abs}} = \left| M(s, t_{\text{male}}) - M(s, t_{\text{female}}) \right|. \quad (1)$$

To facilitate comparability across different metrics, we define the *normalized difference* as:

$$\Delta_{\text{norm}}(\%) = \frac{\Delta_{\text{abs}}}{R} \times 100\,, \quad (2)$$

where $R = \max(M) - \min(M)$ is the *range* of the metric and $I = [\,\min(M), \max(M)\,]$ denotes the corresponding interval of observed values. Since in the general case the theoretical bounds of M may be unknown, we instead rely on empirical estimates derived from our evaluation results, using as $\min(M)$ and $\max(M)$ the smallest and largest values observed for each metric on the challenge set. This approach ensures that all metrics are interpreted within a consistent, data-driven scale.

5.2 Results

Table 3 summarizes the performance of all evaluated metrics on our challenge set, reporting both absolute and normalized differences alongside their empirical ranges and intervals. A paired t-test[8] was conducted

[8] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

Metric	$\Delta_{\text{norm}}(\%)$	Δ_{abs}	R	I
UvA-MT[5]	104.29	0.1314	0.126	[-0.521, -0.395]
rankedCOMET	71.09	0.0455	0.064	[0.468, 0.532]
MetricX-25-QE	11.93	0.5711	4.787	[-8.253, -3.467]
*sentinel-cand	9.70	0.0559	0.576	[-0.237, 0.339]
MetricX-25	7.45	0.6151	8.260	[-12.494, -4.234]
baseCOMET	4.39	0.0058	0.132	[0.435, 0.567]
*COMETKiwi22	4.13	0.0100	0.242	[0.623, 0.865]
Polycand-1	3.72	0.6023	16.184	[76.655, 92.839]
Polycand-2	3.36	0.8590	25.549	[68.230, 93.779]
Polyic-3	3.10	0.6197	19.985	[74.547, 94.533]

Table 3: Absolute and normalized differences for each metric, along with their empirical range R and interval I on the challenge set. Δ_{abs} and Δ_{norm} are reported as averages over all source sentences and language pairs. Baseline metrics are indicated with an asterisk (*).

for each metric and language, showing that all differences between $M(s, t_{\text{male}})$ and $M(s, t_{\text{female}})$ were statistically significant in every case ($p < 0.05$).

All evaluated metrics show a measurable difference between masculine and feminine translations. Notably, the magnitude of these differences, both in absolute and normalized terms, varies considerably across metrics. For example, UvA-MT and rankedCOMET display the largest normalized differences (above 70%), suggesting higher sensitivity to gender variation, whereas metrics such as Polyic-3 and Polycand-2 register much smaller relative changes. While, the empirical ranges R and intervals I differ widely among metrics, reflecting substantial variation in score scales and distributional properties, normalization allows us to fairly compare the different metrics.

Analysis per Language. Table 4 presents the average normalized differences (± standard deviation) across all evaluation metrics for each source–target pair. Across most targets, English as the source leads to higher average differences than Finnish or Turkish. In certain cases, such as Arabic, Czech, Icelandic, Italian, Russian, Serbian, and Ukrainian, the source language changes the difference score substantially (>8%). In contrast, for Greek, Spanish, French, and Portuguese, the results are more stable regardless of the source. This suggests a systematic source-language effect. One possible explanation is that English, as a natural-gendered language, exhibits more extensive gender marking than Finnish or Turkish, which are largely genderless languages, thereby increasing the likelihood that occupational gender associations can be stronger and leak into the evaluation metrics. However, this does not explain why this is not true for all target languages, so further investigation is needed to disambiguate these findings.

When focusing on the target languages, the ranking is stable: Arabic, Russian, and Icelandic yield the highest values; Czech, Italian, Ukrainian, and Serbian follow; French, Greek, Spanish, and Portuguese are consistently lower. This stability across different sources points to target-language characteristics as a dominant factor. Standard deviations parallel the ranking of the averages—larger where averages are higher (especially with English as the source) and smaller where they are lower—indicating greater item-level variability when differences are larger. In high-difference pairs, this variability likely reflects greater diversity in how specific items respond to the different evaluation metrics. Conversely, in low-difference pairs, more uniform behavior is observed across items. Overall, the analysis highlights two main points: while the inherent characteristics of the target language play a dominant role in determining the magnitude of differences, selecting English as the source systematically amplifies these effects, leading to both higher average values and greater variability across samples.

Target \ Source	English	Finnish	Turkish
Arabic	$25.39_{\pm 40.28}$	$16.82_{\pm 32.45}$	$18.09_{\pm 35.52}$
Czech	$22.57_{\pm 34.42}$	$13.38_{\pm 21.14}$	$13.76_{\pm 23.01}$
Greek	$11.55_{\pm 18.48}$	$10.89_{\pm 18.55}$	$10.97_{\pm 19.07}$
Spanish	$11.79_{\pm 19.68}$	$12.32_{\pm 22.05}$	$12.32_{\pm 22.99}$
French	$9.47_{\pm 14.47}$	$8.95_{\pm 13.77}$	$9.13_{\pm 14.60}$
Icelandic	$26.10_{\pm 32.77}$	$16.62_{\pm 19.58}$	$16.55_{\pm 20.51}$
Italian	$22.14_{\pm 35.30}$	$12.59_{\pm 21.41}$	$12.27_{\pm 21.42}$
Portuguese	$11.62_{\pm 20.12}$	$11.41_{\pm 20.14}$	$11.58_{\pm 21.51}$
Russian	$25.63_{\pm 38.98}$	$16.38_{\pm 29.50}$	$15.49_{\pm 27.28}$
Serbian	$20.17_{\pm 33.76}$	$11.64_{\pm 20.18}$	$11.68_{\pm 21.05}$
Ukrainian	$21.64_{\pm 34.86}$	$12.93_{\pm 22.08}$	$12.90_{\pm 21.90}$

Table 4: Average $\Delta_{\text{norm}}(\%)$ with standard deviation for each source-target pair across all models.

[5]Results only with English as the source language.

Occupation	sentinel-cand	COMETKiwi22	MetricX-25	UvA-MT	rankedCOMET	Polycand-2	Polyic-3	MetricX-25-QE	baseCOMET	Polycand-1	AVERAGE
					$\overline{M(s,t_{male})} > \overline{M(s,t_{female})}$						
⚓🔧	+9.7%	+11.9%	+13.1%	+58.0%	+6.1%	+13.6%	+15.5%	+27.0%	+5.9%	+18.8%	+18.0%
💻⚙️	+15.0%	+8.8%	+12.5%	+34.4%	+14.3%	+16.2%	+18.1%	+25.4%	+11.9%	+16.9%	+17.3%
💧🔧	+9.4%	+9.3%	+22.5%	+39.6%	+6.4%	+11.1%	+11.7%	+33.9%	+5.3%	+15.1%	+16.4%
💥💣	+11.5%	+10.3%	+19.9%	+26.7%	+5.6%	+13.5%	+15.4%	+33.3%	+6.7%	+17.5%	+16.0%
🚗🔧	+11.3%	+10.1%	+15.6%	+39.8%	+7.3%	+13.1%	+13.6%	+25.1%	+6.9%	+17.0%	+16.0%
❄️⚙️	+9.3%	+8.1%	+16.7%	+36.4%	+7.5%	+14.6%	+13.8%	+29.5%	+7.2%	+16.5%	+16.0%
⚡⚙️	+13.5%	+11.5%	+13.2%	+31.3%	+9.8%	+12.7%	+14.2%	+24.7%	+10.1%	+15.4%	+15.7%
🎯🐗	+6.4%	+8.5%	+15.8%	+41.8%	+5.1%	+13.0%	+11.4%	+36.0%	+4.8%	+12.9%	+15.6%
⚙️🔨	+9.5%	+10.2%	+17.9%	+34.0%	+4.4%	+13.5%	+15.2%	+28.6%	+5.2%	+16.7%	+15.5%
⚡💻	+15.9%	+9.2%	+11.4%	+24.6%	+10.6%	+15.3%	+21.1%	+20.6%	+9.2%	+15.9%	+15.4%
					$\overline{M(s,t_{male})} < \overline{M(s,t_{female})}$						
👗👔	-2.5%	-6.1%	-13.8%	-22.4%	+1.8%	+0.7%	-3.1%	-25.2%	+0.8%	-2.7%	-7.2%
👧👶	-3.5%	-3.9%	-13.7%	-24.3%	-0.8%	-1.0%	+0.4%	-21.7%	-0.7%	+0.6%	-6.9%
👶🩺	-1.0%	-5.8%	-18.2%	-5.3%	+4.0%	+1.7%	+0.5%	-28.1%	+3.3%	+1.7%	-4.7%
👶🩺+	-2.1%	-2.6%	-21.3%	-4.0%	+6.0%	+1.6%	-0.6%	-28.1%	+5.0%	+4.2%	-4.2%
💊🩺	-1.4%	-2.3%	-10.6%	-8.7%	-0.2%	+1.8%	+2.6%	-15.2%	-0.3%	+1.3%	-3.3%
💊🩺+	-1.2%	-3.8%	-7.5%	-14.8%	+0.6%	+1.9%	+2.3%	-10.5%	+0.1%	+2.1%	-3.1%
🏠🧹	-0.1%	+1.8%	-11.4%	-8.7%	+1.9%	+0.9%	+0.1%	-14.7%	+1.2%	+1.5%	-2.7%
☎️🖊️	-1.1%	+0.8%	-5.7%	-2.3%	+1.1%	+1.3%	+0.8%	-11.4%	+0.9%	+1.8%	-1.4%
🧹🏠	+0.1%	+1.2%	-10.3%	-1.9%	+3.9%	+2.6%	+2.4%	-14.1%	+3.4%	+2.1%	-1.1%
📝🖊️	-0.6%	+1.1%	-7.5%	-5.9%	+3.8%	+3.2%	+4.7%	-13.1%	+3.3%	+3.3%	-0.8%

Table 5: $\Delta_{\text{norm}}(\%)$ per occupation across all evaluated metrics. The occupations referenced in this table are listed in Table 7 along with the respective ISCO codes.

Analysis per Occupation. For this analysis, it is necessary to define not only the *strength* of the effect, captured by the absolute magnitude $\Delta_{\text{norm}}(\%)$, but also its *direction*. Specifically, a "+" is assigned when $\overline{M(s,t_{\text{male}})} > \overline{M(s,t_{\text{female}})}$ (where the overline denotes the mean value), indicating that the metric tends to favor the masculine form, whereas a "–" is assigned when $\overline{M(s,t_{\text{male}})} < \overline{M(s,t_{\text{female}})}$, indicating a tendency to favor the feminine form.

Table 5 presents the top ten occupations with the largest positive and negative normalized differences $\Delta_{\text{norm}}(\%)$ across all evaluated metrics, ranked by their average value across all languages. Occupations are identified using their corresponding ISCO-08 codes, providing a standardized reference for each job category.

The occupations appearing in the positive segment of the table (e.g., ISCO 3151 "Ships' Engineers", ISCO 7126 "Plumbers and Pipe Fitters", ISCO 7542 "Shotfirers and Blasters") are predominantly roles that are culturally and historically associated with men and are often perceived as physically demanding or male-dominated. In contrast, the negative segment features occupations such as ISCO 5241 "Fashion and Other Models", ISCO 5311 "Child Care Workers", and ISCO 3222 "Midwifery Professionals", which are stereotypically viewed as female-oriented professions. This pattern suggests that the evaluated metrics may reproduce gender stereotypes in their outputs, ultimately reinforcing such biases in MT systems.

It is also noteworthy that the magnitude of the positive differences is, in absolute terms, generally greater than that of the negative differences. For example, the tenth-highest positive difference toward the masculine form substantially exceeds the highest negative difference toward the feminine form. This asymmetry indicates an overall tendency of the evaluated metrics to favor masculine forms in translation.

6 Gender Density Analysis

Different human-evaluation approaches of translation penalize errors (such as using the wrong gender) differently: direct assessment applies a conceptual approach, penalizing each error only once (Graham et al., 2017), whereas MQM (Lommel et al., 2013) penalizes every occurrence of the error in a text. To better understand how QE metrics capture gender related aspects, we analyse the correlation between the

Metric	English	Finnish	Turkish
*COMETKiwi22	0.17	0.21	0.22
MetricX-25	0.02	0.09	0.08
MetricX-25-QE	0.06	0.17	0.19
Polycand-1	0.18	0.22	0.22
Polycand-2	0.17	0.21	0.20
Polyic-3	0.12	0.16	0.18
UvA-MT	0.47	—	—
baseCOMET	0.20	0.21	0.22
rankedCOMET	0.09	0.10	0.11
*sentinel-cand	0.02	0.02	0.02

Table 6: Pearson correlation coefficients between gender density and metrics' bias. All metrics show a positive correlation and all correlations are statistically significant ($p < 0.05$). Baseline metrics are indicated with an asterisk (*).

bias metric (normalized difference score $\Delta_{\text{norm}}(\%)$) and gender density in the text. We define *gender density* as the normalized count of gendered words in the text. Since the two texts (t_{male}, t_{female}) are identical except for the gendered words, we calculate the gender density as the proportion of differing words to the average text length (in words):

$$\text{gender density} = \frac{\text{worddiff}(t_{male}, t_{female})}{\text{average}(\text{len}(t_{male}), \text{len}(t_{female}))}$$

As presented in Table 6, we observe a positive correlation between metric bias and gender density for all metrics with varying correlation strength, i.e. *texts with a higher number of gendered words result in more biased metric evaluations*. Among the evaluated metrics, Polycand metrics show the strongest correlations overall to gender density. COMET-based metrics (COMETKiwi22, baseCOMET, rankedCOMET) also show strong correlations, while MetricX-25 and MetricX-25-QE consistently show weak correlations. Sentinel-cand metric shows minimal correlations, suggesting it may be less sensitive to gender density. Detailed correlation plots for each metric are in appendix C. The findings suggest that the metrics apply a cumulative preference and penalty for gender in a text.

7 Conclusions

We presented GAMBIT+, a large-scale, fully parallel challenge set containing paired masculine and feminine translations of gender-ambiguous occupations across 33 source–target language pairs, and used it to benchmark 10 QE metrics. Our evaluation revealed consistent, statistically significant score shifts driven solely by grammatical gender—often more pronounced in certain language pairs and occupations—with a clear overall tendency to favor masculine forms. These findings indicate that current QE metrics are not gender-fair and should be systematically audited and calibrated.

Future work will broaden the scope of our analysis to include a wider range of QE metrics, with a focus on identifying specific characteristics that make them more susceptible to gender bias. We also plan to investigate the interplay between MT systems and QE metrics, exploring how system outputs and metric evaluations align or diverge in terms of gender bias. We also plan to extend GAMBIT+ with translations of varying quality to investigate whether translation quality influences the biases of QE metrics—for example, whether poor translations make gender distinctions more or less apparent due to incorrect gender agreements across the sentence. Finally, we aim to extend the use of GAMBIT+ beyond MT, testing its applicability and value in evaluating gender bias in other natural language generation tasks.

Acknowledgments

This work is part of the UTTER project, supported by the European Union's Horizon Europe research and innovation programme via grant agreement 101070631, and by the FCT project "OptiGov", ref. 2024.07385.IACDC (DOI 10.54499/2024.07385.IACDC), funded by the PRR under the measure RE-C05-i08.m04. Giuseppe Attanasio and Chrysoula Zerva are also supported by the Portuguese Recovery and Resilience Plan through projects C645008882-00000055 (Center for Responsible AI) and UID/50008: Instituto de Telecomunicações. Chrysoula Zerva is also supported by an unrestricted gift from Google (Google Research Scholar). Wafaa Mohammed acknowledges travel support from ELIAS (GA no 101120237).

References

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in danger? evaluating speech translation technology on the MuST-SHE corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin

Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Anna Farkas and Renáta Németh. 2022. How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences & Humanities Open*, 5(1):100239.

Mingqi Gao and Xiaojun Wan. 2022. Social biases in automatic evaluation metrics for nlg. *arXiv preprint arXiv:2210.08859*.

Sarthak Garg, Mozhdeh Gheini, Clara Emmanuel, Tatiana Likhomanenko, Qin Gao, and Matthias Paulik. 2024. Generating gender alternatives in machine translation. *arXiv preprint arXiv:2407.20438*.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 901–912.

Eleni Gkovedarou, Joke Daems, and Luna De Bruyne. 2025. Gender bias in english-to-greek machine translation. *3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*.

Atmika Gorti, Aman Chadha, and Manas Gaur. 2024. Unboxing occupational bias: Debiasing llms with us labor data. In *Proceedings of the AAAI Symposium Series*, volume 4, pages 48–55.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Aida Kostikova, Joke Daems, and Todor Lazarov. 2023. How adaptive is adaptive machine translation, really? a gender-neutral language use case. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 95–97, Tampere, Finland. European Association for Machine Translation.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*, pages 12–24.

Manuel Lardelli, Giuseppe Attanasio, and Anne Lauscher. 2024. Building bridges: A dataset for evaluating gender-fair machine translation into German. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7542–7550, Bangkok, Thailand. Association for Computational Linguistics.

Manuel Lardelli and Dagmar Gromann. 2023. Gender-fair post-editing: A case study beyond the binary. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 251–260, Tampere, Finland. European Association for Machine Translation.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi kiu Lo, Vilém Zouhar, Frédéric Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David I. Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Seventh Conference on Machine Translation (WMT 2025)*, Miami, Florida, USA. Association for Computational Linguistics.

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Sujal Maharjan and Astha Shrestha. 2025. Rankedcomet: Elevating a 2022 baseline to a top-5 finish in the wmt 2025 qe task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms. *arXiv preprint arXiv:2503.04372*.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Eva Tsouparopoulou, Dimitris Parsanoglou, Maria Symeonaki, and Giorgos Stamou. 2024. Gost-mt: A knowledge graph for occupation-related gender biases in machine translation. *arXiv preprint arXiv:2409.10989*.

Orfeas Menis-Mastromichalakis, George Filandrianos, Maria Symeonaki, Glykeria Stamatopoulou, Dimitris Parsanoglou, and Giorgos Stamou. 2025. Gender bias in machine learning: insights from official labour statistics and textual analysis. *Quality & Quantity*, pages 1–35.

Angela Balducci Paolucci, Manuel Lardelli, and Dagmar Gromann. 2023. Gender-fair language in translation: A case study. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 13–23, Tampere, Finland. European Association for Machine Translation.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Silvia Alma Piazzolla, Beatrice Savoldi, and Luisa Bentivogli. 2023. Good, but not always fair: An evaluation of gender bias for three commercial machine translation systems. *arXiv preprint arXiv:2306.05882*.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023a. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023b. Hi guys or hi folks? benchmarking gender-neutral machine translation with the GeNTE corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140, Singapore. Association for Computational Linguistics.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8358–8375, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Argentina Anna Rescigno, Eva Vanmassenhove, Johanna Monti, Andy Way, and 1 others. 2020. A case study of natural gender phenomena in translation-a comparison of google translate, bing microsoft translator and deepl for english to italian, french and spanish. In *CEUR Workshop Proceedings*, pages 359–364. AILC-Associazione Italiana di Linguistica Computazionale.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. *arXiv preprint arXiv:2410.00545*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuanjing Huang. 2022. BERTScore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Eva Vanmassenhove. 2024. Gender bias in machine translation and the era of large language models. *Gendered Technology in Translation and Interpreting: Centering Rights in the Development of Language Technology*, page 225.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Di Wu and Christof Monz. 2025. Uva-mt at wmt25 evaluation task: Llm uncertainty as a proxy for translation quality. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and André FT Martins. 2024. Watching the watchers: Exposing gender disparities in machine translation quality estimation. *arXiv preprint arXiv:2410.10995*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. Cometpoly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

A Prompts

Below is the prompt designed for the translation task. In this context, the variable {gender} can take the value "male" or "female," {occupation} refers to the specific occupation under study as defined by the GAMBIT framework, and {lang} denotes the target language. The prompt instructs the translation model to produce a faithful and semantically equivalent rendering of the source text, adapting all gendered references to match the specified {gender} form for the given {occupation}.

Gendered Translation Prompt

You are an expert professional translator. Translate the following English text into {lang}.

Important instructions:

- Any reference to the word "{occupation}" must be translated using the {gender} form in {lang}.

- Be extremely careful to preserve the exact meaning, tone, and semantics of the original text.

- Do **not** modify, omit, or add any information other than adapting the gender reference to {gender} for the {occupation}.

- Return **only the translation**, without any additional comments, notes, or explanations.

Text:

{text}

Translation:

Below is the prompt used for the gender-adaptation procedure in translations. In this setting, the variable {gender} assumes the opposite value from that assigned in the initial stage of the workflow-that is, if the initial assignment is "male," it is here set to "female," and vice versa. The variable {occupation} corresponds to the occupation label provided in the GAMBIT dataset, while {lang} specifies the language into which the translation is to be rendered. The instructions require the translation model to maintain complete semantic fidelity to the source text, altering only the gendered form of the {occupation} in accordance with the designated {gender}, without introducing, removing, or rephrasing any other content.

Gender Adaptation Prompt

You are an expert professional translator. Translate the following English text into {lang}.

Important instructions:

- Any reference to the word "occupation" must be translated using the {gender} form in {lang}.

- Be extremely careful to preserve the exact meaning, tone, and semantics of the original text.

- Do ***not*** modify, omit, or add any information other than adapting the gender reference to {gender} for the "{occupation}".

- Return ***only the translation***, without any additional comments, notes, or explanations.

Text:

{text}

Translation:

For languages without grammatical gender, namely Turkish and Finnish, no gender adaptation process is applied. In these cases, only the standard translation prompt is employed, as provided below, ensuring that occupational and role references remain neutral in accordance with the structural characteristics of the target language.

Genderless Translation Prompt

You are an expert professional translator. Translate the following English text into lang.

Important instructions: - Pay close attention to the gender of any occupations or role titles. - If the gender is not clear or not mentioned, do not assume or infer it; keep the translation gender-neutral or ambiguous as appropriate in the target language. - Return only the translation, without any additional comments, notes, or explanations.

Text:

{text}

Translation:

For the evaluation phase, the following prompt was employed to compare two texts in the target language. Its primary function was to determine whether any differences between the two texts were solely attributable to the explicit mention of the gender of the specified occupation or whether other differences

were present. Text 1 and Text 2 correspond to the two translations of the same source text, each rendered with a different specified gender for the occupation under study.

Evaluation Prompt

You are a linguistic comparison assistant. Analyze the following two {lang} texts and identify the differences between them.

If the only differences are related to the explicit mention of the gender of the occupation occupation_title (e.g., masculine vs. feminine forms of the occupation {occupation_title}), respond with "yes".

If you find any other type of difference, respond with "no" and list all the differences you found.

Return only the answer, without any additional comments, notes, or explanations, especially when the answer is "yes".

Text 1:

{text1}

Text 2:

{text2}

B ISCO Code–Occupation Mapping

Table 7 provides the mapping between the ISCO codes referenced in Table 1 and their corresponding occupational titles as defined in the ISCO-08 classification.

ISCO Code	ISCO Name	Emojis
3151	Ships' Engineers	⚓🔧
3513	Computer Network and Systems Technicians	💻⚙️
7126	Plumbers and Pipe Fitters	💧🔧
7542	Shotfirers and Blasters	💥💣
7231	Motor Vehicle Mechanics and Repairers	🚗🔧
7127	Air Conditioning and Refrigeration Mechanics	❄️⚙️
7421	Electronics Mechanics and Servicers	⚡⚙️
6224	Hunters and Trappers	🎯🐗
7213	Sheet Metal Workers	⚙️🔨
3114	Electronics Engineering Technicians	⚡💻
5241	Fashion and Other Models	👗👔
5311	Child Care Workers	🧒👶
2222	Midwifery Professionals	👶🩺
3222	Midwifery Associate Professionals	👶🩺+
2221	Nursing Professionals	💊🩺
3221	Nursing Associate Professionals	💊🩺+
5151	Cleaning and Housekeeping Supervisors	🏠🧹
4226	Receptionists (general)	☎️🖊️
9111	Domestic Cleaners and Helpers	🧹🏠
4120	Secretaries (general)	📝🖊️

Table 7: Mapping of emojis from Table 5 to their corresponding occupational titles and the respective ISCO codes.

C Gender-Density Correlation Plots

Figure 1 shows the correlation plots of gender density to normalized score difference for each metric. Plots are shown for English as a source language. Other source languages show the same trend.

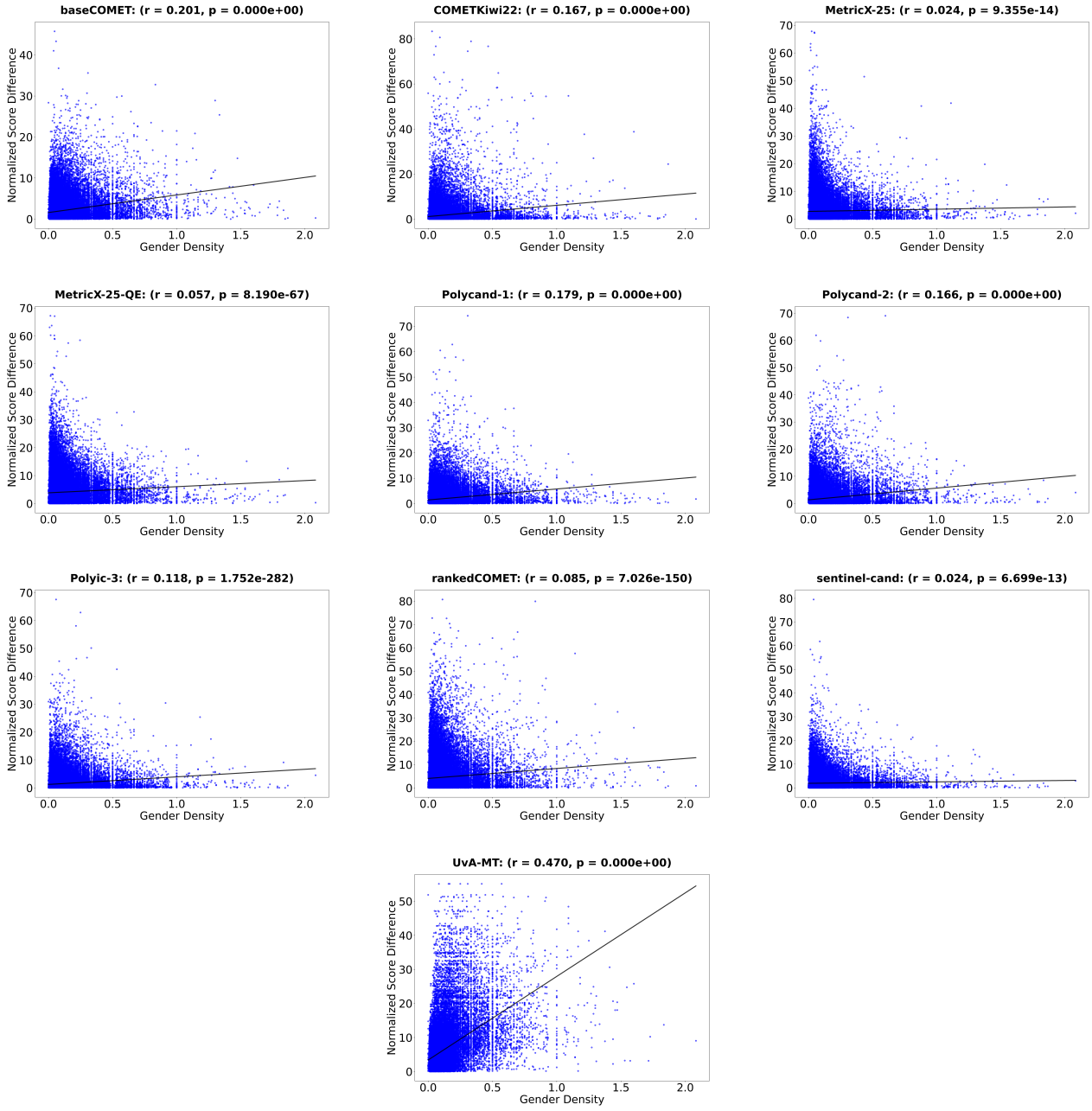


Figure 1: Gender density to normalized score difference correlation for all metrics.

Implementing and Evaluating Multi-source Retrieval-Augmented Translation

Tommi Nieminen and **Jörg Tiedemann** and **Sami Virpioja**
University of Helsinki
{tommi.nieminen,jorg.tiedemann,sami.virpioja}@helsinki.fi

Abstract

In recent years, neural machine translation (NMT) systems have been integrated with external databases with the aim of improving machine translation (MT) quality and enforcing domain-specific terminology and other conventions in the MT output. Most of the work in incorporating external knowledge with NMT has concentrated on integrating a single source of information, usually either a terminology database or a translation memory. However, in real-life translation scenarios, all relevant knowledge sources should be used in parallel. In this article, we evaluate different methods of integrating external knowledge from multiple sources in a single NMT system. In addition to training single models trained to utilize multiple kinds of information, we also ensemble models that have been trained to utilize a single type of information. We evaluate our models against state-of-the-art LLMs using an extensive purpose-built English to Finnish test suite.

1 Introduction

Most NMT systems receive as their input a source sentence on its own, without any additional context. This is problematic, as producing a correct translation often requires information that is external to the source sentence. For instance, source sentences that are translated as part of a larger document have to be consistent with other parts of the document. Even if the translation of a sentence is not constrained by document context, the translation often needs to conform to terminological or phraseological conventions of a genre, domain, or a house style. Beyond acting as a contextual constraint, external information may also improve translation quality by providing the NMT system with translation examples. These examples can simplify the task of translation, as the NMT system does not have to generate the translation from scratch, but can adapt the provided external information.

One method of providing relevant external information to an NMT system is to query an external database for data based on the source sentence, and then either include the retrieved information as part of the NMT system input or constrain the decoding process based on the retrieved information. As similar information retrieval approaches used with large language models are called retrieval-augmented generation (RAG) (Lewis et al., 2020), we will refer to this family of methods as retrieval-augmented translation (RAT), following Hoang et al. (2023).

Even though almost all published RAT methods concentrate on a single kind of retrieved information (usually either terminology or translation memory matches), in actual practical translation scenarios all the kinds of retrieved information are used simultaneously. For instance, a human translator working in a computer-assisted translation (CAT) tool will be provided with matches from both terminology database and translation memories, and they will need to make their translations conform with both of these information sources.

In this article, we introduce several NMT systems, which can utilize different kinds of retrieved information when generating translations. We experiment with both single NMT models that are trained to utilize multiple kinds of information, and with systems that combine models that have been trained to utilize a single kind of information using a novel ensembling method called contrastive ensembling. We also compare our models with instruction-tuned LLMs, which have a native capability of utilizing multiple types of retrieved information.

Our models are trained to utilize two kinds of information:

1. **Fuzzy matches**: Parallel sentences retrieved from a translation database (translation memory or TM). The search is based on the edit distance between the source sentence being trans-

Proceedings of the Tenth Conference on Machine Translation, pages 327–339
November 8-9, 2025

lated and the source sentence in the translation database. Usually the matches are restricted to those whose normalized edit distance exceeds a specific threshold (in the translation industry a threshold corresponding to a similarity level of 70 percent is normally used).

2. **Term matches**: Terms retrieved from a terminology database (termbase or TB). The database is searched for all the sub-strings of the source sentence being translated, and all terms where the source term matches the sub-string are returned. As the TB usually contains the terms in their dictionary forms, the sub-strings are lemmatized or stemmed before the search.

The motivation for including these two kinds of information is that they are routinely used in the CAT tools that professional translators use. They represent the types of information that are readily available and have been found useful in real-life translation workflows (Hutchins, 1998). Building the RAT system around widely used types of information also ensures that it can easily be integrated into existing workflows.

As mentioned, RAT can be used in two ways: either as constraining the MT system to utilize the retrieved information in its output (especially in the case of terminology), or as providing contextual information to enable the generation of better translations. In this article, we are mainly concerned with constraining RAT, as it has applications in professional translation. One of the problems in developing constraining RAT systems is that the common MT evaluation methods, such as BLEU and COMET, are not well suited to evaluating them, as they provide no information on how well the retrieved information has been utilized. To help us develop and evaluate our models, we therefore compiled an extensive test suite, which contains test cases consisting of source sentences, terms, fuzzy matches, and tests that can be used to check whether the terms and fuzzy matches are used in translations.

2 Related work

We structure our system around retrieval methods that have been used in professional translation since the 1960s. These methods have been developed gradually and organically within the translation industry, so their origins are often unclear. For background on fuzzy match and term retrieval, see Hutchins (1998).

The first MT method that can be characterized as RAT was example-based machine translation (EBMT) (Nagao, 1984), where translations were generated based on examples retrieved from a translation database. Statistical machine translation (SMT) methods could also be characterized as a form of RAT, as they rely on retrieving partial translations from a database of translation fragments, and retrieval was also more explicitly integrated into SMT systems (Koehn and Senellart, 2010).

In the context of NMT, one of the first methods recognizable as RAT was introduced in Gu et al. (2017), where fuzzy matches were retrieved from a TM and the attention component of the model was modified to cover the matches in addition to the source sentence. Constraining NMT to adhere to retrieved terminology was first introduced in Hokamp and Liu (2017), where the beam search decoder is modified to always produce the specified target terms in the output. Song et al. (2019) was the first to implement RAT using data-based methods, by replacing sub-strings in the source sentences of the training data with equivalent target language sub-strings. Dinu et al. (2019) introduced data-based RAT for terminology and Bulte and Tezcan (2019) for fuzzy matches.

While most work on RAT has concentrated on a single kind of information, there has been some recent work on unified RAT, where NMT systems can utilize multiple kinds of information. Wang et al. (2023) prefix the inputs and outputs of their model with three different kinds of retrieved information: fuzzy matches, translation templates, and terms. Raunak et al. (2024) fine-tune an NMT model with data augmented with many different kinds of instructions, some of which are similar to RAT, such as an instruction to utilize a particular term in the translation. Moslem et al. (2023) experiment with prompting an LLM with both terms and fuzzy matches to generate adapted translations. What differentiates our approach from the previous implementations is that we focus exclusively on the types of information that are routinely used in CAT tools, which enables easy integration with professional workflows. We also use an ensemble of specialized models in addition to a single model trained to utilize multiple types of information.

The concept of generating translations using multiple different inputs originates from the field

of multi-source translation (Och and Ney, 2001), where the different inputs are equivalent source sentences in different languages. Firat et al. (2016) were the first to implement multi-source translation by ensembling different NMT models that are provided with different inputs, which resembles our ensembling method.

For evaluation, we will use a dedicated test suite for the phenomena we want to tackle. Test suites have been used for evaluating MT quality since at least the 1990s (King and Falkedal, 1990), and they have become more popular in recent years (see for example Macketanz et al., 2022), as the dramatically improving MT quality has created A demand for more granular evaluation methods. As far as we know, our test suite is the first suite designed specifically for RAT evaluation.

3 Models

We train a selection of models, including separate term and fuzzy match RAT models, and unified RAT models, which process both types of retrieved information. For term models we train models that can process a single term, and models which can process up to ten terms. For fuzzy models, we train models that can process a single fuzzy match, and models that can process up to three fuzzy matches.

Models are created with continued training using the Tatoeba-Challenge (Tiedemann, 2020) models as the base models. As our test suite is English to Finnish, we only train models in that language direction. For all models but one we use the standard transformer model *opusTCv20210807+bt-2021-09-01* as the base model. To see the effect of model size, we also train one model using a base model with the transformer-big architecture (*opusTCv20210807+news+bt_transformer-big_2023-04-13*). Continued training has many advantages compared to training the models from scratch: as continued training is much faster, it is easier to test different model variations and the carbon footprint of the training is smaller. The base models can also be used as strong baselines for evaluation, as they have been trained on all the available data from the OPUS corpus (Tiedemann, 2009).

Our models use special symbols to separate the retrieved information from normal source text (see the left column in Figure 1 for an example of how the symbols are used). The vocabularies of the base models do not have any spare symbols that can be used as these special symbols, so we need to re-purpose some of the existing symbols. We pick ten of the least common symbols from the vocabulary, and assign them as our special symbols. Not all symbols are used in the experiments, but we reserve extra symbols in case more are needed in future experiments with the same models. As the vocabularies remain otherwise identical, we can easily ensemble the trained models with each other and with the base model.

Training is continued with a high-quality subset of the Tatoeba-Challenge dataset that was originally used to train the base models. Ten million sentences are included in the continued training subset. The subset does not include data from crawled corpora due to quality problems associated with them (Kreutzer et al., 2022). The data is also scored with BiCleaner-AI (Zaragoza-Bernabeu et al., 2022), and sentence pairs scoring less than 0.7 are excluded from the subset. The duration of continued training is one epoch, and the learning rate is set to 0.00001 to prevent catastrophic forgetting (McCloskey and Cohen, 1989).

For both fuzzy and term models, the training set is annotated with the appropriate RAT data for that type (see Figure 1 for examples of the annotations). The models are trained using the Marian NMT framework (Junczys-Dowmunt et al., 2018).

3.1 Term models

The terminology models are trained using the data augmentation method first introduced in Dinu et al. (2019): source terms are identified in the source sentence, and target terms are appended to the source sentence after the corresponding source terms. Following Bergmanis and Pinnis (2021) we append the source sentence with lemma forms of the target terms instead of the surface forms, in order to train the model to inflect the provided target terms on the target side instead of copying them directly. We also follow Bergmanis and Pinnis (2021) in using synthetic terms, which are generated by aligning the parallel data on the token-level using *fast-align* (Dyer et al., 2013) and then selecting aligned noun and verb phrases as the synthetic terms. *Stanza* (Qi et al., 2020) was used for lemmatization and to identify noun and verb phrases.

3.2 Fuzzy match models

The fuzzy match models are trained using the Neural Fuzzy Repair (NFR) method introduced in Bulte and Tezcan (2019). We use the continued training

subset as a translation database from which fuzzies are retrieved. The database is searched for matches using the *fuzzy-match*[1] library, and the target sides of the matches are prefixed to the source sentences to produce the training data.

Preparing training data for fuzzy match models is more complicated than for term models. In the term model training data we can always make sure that the target term appended to the source sentence is actually present on the target side, but the situation is different with fuzzy matches. Fuzzy matches have to be retrieved from naturally occurring data, as producing them synthetically is not feasible. Also, fuzzy matches are not binding in the sense that terminology is: in the case of terms, the target term can almost always be used in a translation, but it is very common to have a fuzzy match that cannot be used in any valid translation for a source sentence. Because of this, the ideal training data for fuzzy match models consists of the following types of sentence pairs:

1. **Positive examples**: Sentence pairs, where the source sentence is appended with a usable fuzzy match (i.e. a fuzzy match that can be used in a valid translation for the source sentence), and parts of that fuzzy match are present in the target sentence.

2. **Negative examples**: Sentence pairs, where the source sentence is appended with an unusable fuzzy match (i.e. a fuzzy match that cannot be used in a valid translation for the source sentence), and that fuzzy match is not used in the target sentence.

If fuzzy matches are retrieved based on source similarity, the mix of training examples is not optimal, as it will contain many examples where a usable fuzzy match is not used on the target side. On the other hand, if fuzzy matches are retrieved based only on target similarity, the training set will only contain positive examples, and the model will learn to always copy from the fuzzy matches, even when inappropriate. Because of this, we retrieve fuzzies using both source and target similarity, as in Nieminen et al. (2025).

3.3 Unified model

In addition to the separate term and fuzzy models we also train a unified model, which is trained on both two types of data. The training data for this model is generated by merging the training data for the term and fuzzy models. Specifically, we combine the training data of those models, that are trained to process multiple terms or fuzzies, as the unified model also has to process multiple terms and fuzzies.

4 Multisource ensembling

In addition to using a unified model capable of processing both terms and fuzzies, we ensemble models trained to utilize a single kind of information to produce a system that can utilize multiple kinds of information. Each model in the ensemble has its own input, which is prefixed (full matches) or interleaved (terms) with appropriate retrieved information. See Figure 1 for a schematic of the inference pipeline. The ensemble decoder is implemented by modifying the generation functionality in the HuggingFace *Transformers* library[2].

We experiment with the following ensembling methods (see Figure 2 for a visual example):

1. **Naive ensembling**: The next token probabilities of each model in the ensemble are averaged during inference, with equal weight given to each model.

2. **Contrastive ensembling**: Naive ensembling dilutes the effect of the individual models in the ensemble. This is undesirable in our use case, as we want certain models to have more effect than other models at certain phases of generation. For instance, when the next token to be generated is part of the translation for a source term, we want to emphasize the effect of the terminology model that has been provided that specific source term as part of its input. To achieve this, we compare the next token probability distribution for each model to the token probabilities of a contrast model which has not been provided any external information. If a symbol's probability of a model differs significantly from its probability with the base model, the weight of the symbol is boosted in the ensemble (see source code[3] for implementation details). We use the base model from which the RAT models have been trained from as the contrast model.

[1] https://github.com/SYSTRAN/fuzzy-match

[2] https://github.com/huggingface/transformers

[3] https://github.com/Helsinki-NLP/OpusDistillery/blob/modularization/pipeline/hf/multisource_eval.py

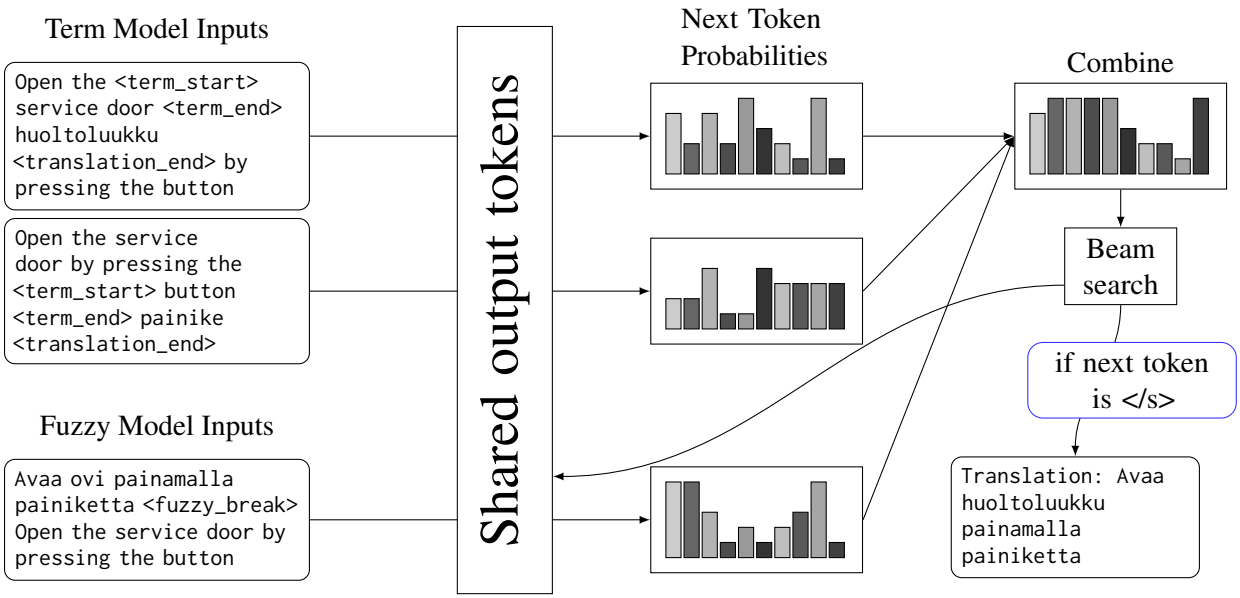


Figure 1: The ensembling inference pipeline with three models and three different inputs. The output tokens are always shared between the models. Note differences in utilization: terms are used completely and inflected, and fuzzy matches are used partially.

Ensembling serves two purposes. First it allows us to utilize multiple kinds of retrieved information during inference, but it also acts as conventional ensembling, the purpose of which is to improve generic output quality. We also hypothesize that ensembling models trained for different RAT methods and provided with different inputs during inference will enhance the quality improving effect of ensembling, as it has been shown (Hoang et al., 2024) that ensembling diverse models produces better results than ensembling similar models, such as different checkpoints of a single training run.

5 Evaluation

RAT systems can be used both for improving generic translation quality and for domain adaptation. When used to improve generic translation quality, it is not relevant whether the translations actually adhere to the terminology and phraseology used in the retrieved examples. For domain adaptation, however, adherence to the retrieved examples is important.

We evaluate model performance from two points of view: correct utilization of the retrieved information in the translations, and general translation quality. For evaluating the correct utilization, we use our test suite, which is covered in detail below. The main purpose of general translation quality evaluation is to see whether continued training with RAT-augmented data degrades generic translation quality. Measuring general translation quality for RAT systems is complicated by the fact that RAT systems are meant to be used with retrieved information: their performance when not provided with any such information is almost irrelevant, as any RAT system can be paired in production with a back-off system that processes source sentences with no retrieved information.

Therefore evaluating RAT systems with standard evaluation sets, which are not paired with retrieved information, does not provide much useful information about the primary use case for RAT systems. Because of these concerns, we use our test suite also as the generic quality test set, so that we can provide retrieved information to the systems. As the test suite does not contain any reference translations, we use the reference-free *wmt23-cometkiwi-da-xl* metric (Rei et al., 2023) for evaluation. This metric has been shown (Freitag et al., 2024) to correlate reasonable well with human judgments and to perform better than many standard reference-based metrics such as BLEU and chrF.

5.1 Test suite

To evaluate the utilization of retrieved information in the generated translations, we created a test suite consisting of source sentences, examples of retrieved information, and tests for checking whether the examples of retrieved information have been utilized in the translations. The test suite was gen-

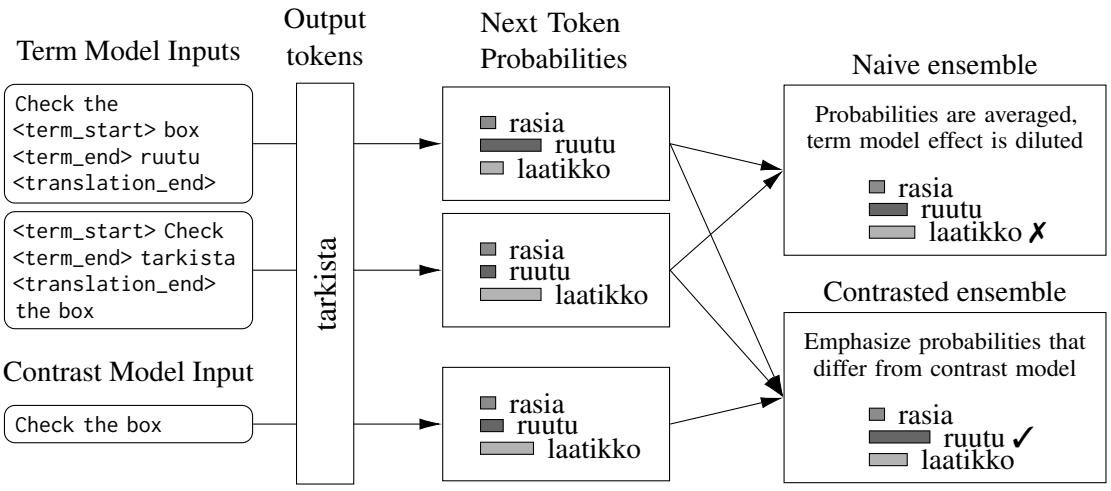


Figure 2: An example of the two ensembling methods. The graph represents a scenario where the source sentence is *Check the box*, and so far the output token *tarkista* has been generated. The word *box* that is being translated in the example is highly polysemous, but the context favours the translation *laatikko*. However, our terminology stipulates that the translation *ruutu* must be used. Naive ensembling produces incorrect output due to treating all models as equal in the context, while contrasted ensembling correctly emphasizes the model that is relevant in the context.

erated semi-automatically: an LLM was used to generate the data, which was then validated and edited by a human reviewer. Initially we tested whether a test suite could be created in a single phase using an LLM (*DeepSeek-V3*), by prompting the LLM to produce complete test cases. While this approach worked for a small test suite (ten or so test cases), when prompted to generate a larger test suite (tens of test cases), the LLM output quality started to degrade noticeably. Because of this degradation, we decided to divide the task into smaller sub-tasks.

As the first step of test suite creation, we prompted the LLM to generate English source sentences. Again, prompting for a large amount of output lead to noticeable output quality degradation, such as repetitive and short sentences. To generate a sufficient amount of high-quality source sentences, we also had to subdivide the sentence generation task to sub-tasks. First we specified seven domains (medical, pharmaceutical, public administration, EU texts, IT administration, IT customer support, and legal), for which sentences could be separately generated for. To add variety, for each of the domains we prompted for the generation of sentences in three different length classes (short, medium, long). In total, we therefore used 21 different prompts, each requesting ten sentences, to generate the source sentences.

The LLM was then prompted to generate a number of fuzzy matches for each of the generated source sentences. For each sentence, three types of fuzzies were generated:

- **Addition fuzzies**: Sentences, which contain additional tokens compared to the source sentence.
- **Deletion fuzzies**: Sentences, which are modifications of the source sentence where some part has been removed.
- **Replacement fuzzies**: Variation of the source sentence where some part has been replaced with a semantically different part.

Note that at this phase only the source side of the fuzzy matches were generated, the target sides were generated later in a separate phase. The edit distance between the generated fuzzies and source sentences was calculated, and all fuzzies below a 70 percent similarity threshold (standard in the translation industry) were automatically discarded.

Once the source sentences and fuzzies had been generated, a human reviewer validated them using a graphical interface (also generated using an LLM) displaying each source sentence and its fuzzy matches on different pages. The reviewer selected 1-5 credible fuzzies for each source sentence. If there were no suitable fuzzies, the reviewer deleted the sentence from the test suite. The reviewer also corrected minor mistakes in some sentences and fuzzy matches.

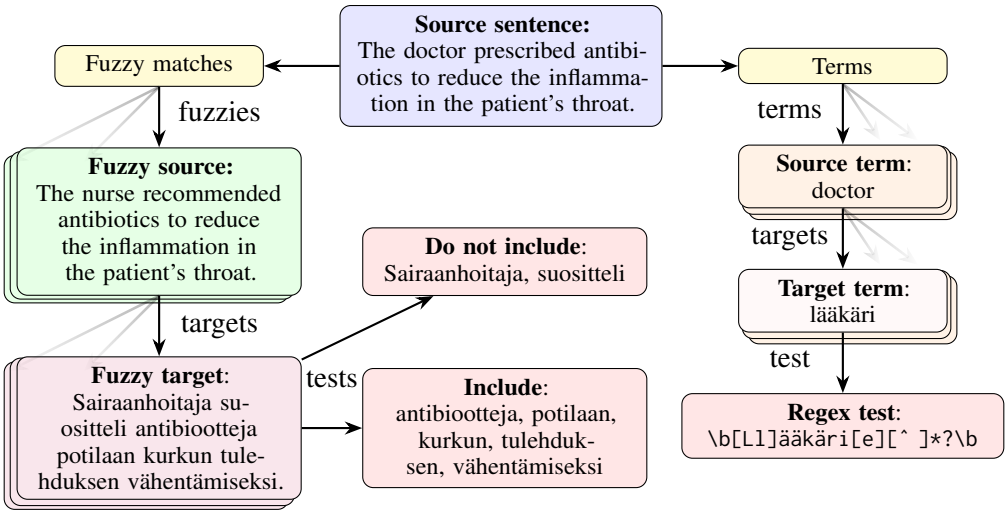


Figure 3: A single test sentence from the test suite. Fading arrows indicate that there can be multiple elements (only one element is shown in the graph to save space).

After the fuzzy validation phase, we prompted the LLM to generate terms and their Finnish translations for each source sentence, as well as regular expression tests for checking whether the term is used in Finnish sentence. The prompt specified that there should be multiple plausible translations for each term. This is important, since if a term has only one plausible translation, it cannot be used to make distinctions between MT systems, as most MT systems are likely to include the correct translation in their output. The terms were also validated manually by a human reviewer using an LLM-generated graphical interface, where the reviewer could select the most plausible terms and their translations, and remove test cases for which no plausible terms had been generated. The reviewer also corrected terms using the interface. The regular expression tests for correct term use that the LLM generated were not usable as such due to them not reflecting Finnish morphology, so the reviewer had to manually correct the tests.

The next phase was to generate translations for the fuzzies. For all other generation phases we used *DeepSeek-V3*, but as the quality of its English to Finnish translations was very uneven, we switched to *GPT4.1* model for generating the translations. The reviewer again reviewed, validated and corrected the translations using an LLM-generated user interface.

The last phase of the test suite generation was generating tests for identifying correct usage of fuzzy matches. Unlike naturally occurring fuzzies, all of the fuzzies in the test suite can be used to construct a valid translation for the source sentence. The test suite is also aimed at scenarios, such as professional translation, where using as much of the fuzzy as possible is desirable, in order to ensure consistency with previous translations. Because of this, we decided to use simple lexical tests to identify whether a fuzzy has been correctly used in a translation. We divided the tokens in each fuzzy into two sets: 1. **Include**: those that correspond semantically to tokens in the source sentence, and should be used in the translation and; 2. **Do not include**: those tokens that have no semantic equivalents in the source sentence and should not be used in the translation.

To create the **Include** and **Do not include** sets, we prompted an LLM with the source sentence, fuzzy source, and the translation of the fuzzy, and instructed the LLM to divide the tokens into the two sets. We also tested traditional word alignment methods, but their accuracy turned out to be too low.

The completed test suite contains 128 source sentences, 434 fuzzy source sentences, 620 fuzzy translations, 403 terms, and 870 term translations. In total, over 200,000 different test cases can be constructed by combining the fuzzy source sentences, fuzzy translations and terms in different combinations. To keep the test suite size manageable for testing (especially with larger models), we create a limited test suite by first organizing the test suite into groups based on how many terms and fuzzy matches each test case contains. Then we randomly pick 50 test cases from each group,

whilst making sure to pick a similar amount of test cases from each domain. We exclude test cases with more than three fuzzy matches to simplify the evaluation task. Our limited test suite contains 1150 test cases.

During evaluation, the test suite produces eight different scores:

1. **Term success (TS)**: the target sentence passes the term test.
2. **Term failure (TF)**: the target sentence fails the term test.
3. **Suitable fuzzy token included (FP)**: the target sentence contains a fuzzy token from the **Include** token list.
4. **Suitable fuzzy token not included (FN)**: the target sentence does not contain a fuzzy token from the **Include** token list.
5. **Invalid fuzzy token not included (IN)**: the target sentence does not contain a token from the **Do not include** token list.
6. **Invalid fuzzy token included (IP)**: the target sentence contains a token from the **Do not include** token list, indicating over-copying.
7. **Suitable fuzzy token bigram included (BP)**: the target sentence contains a bigram of fuzzy tokens from the Include token list.
8. **Suitable fuzzy token bigram not included (BN)**: the target sentence does not contain a bigram of fuzzy tokens from the Include token list.

The bigram scores are included to reward using the same order of tokens in the translation as in the fuzzy. If there are multiple fuzzy matches available to the system, the fuzzy match that has the best overall score is used as the basis of the fuzzy match scores.

We report average term and fuzzy scores for each system, which are calculated with the following formulas:

$$TermScore = \frac{TS}{TS + TF} \quad (1)$$

$$FuzzyScore = \frac{1}{3}\left(\frac{FP}{FP + FN} + \frac{IN}{IN + IP} + \frac{BP}{BP + BN}\right) \quad (2)$$

To facilitate comparison between systems, we also produce a composite score using the following formula:

$$CompositeScore = \frac{5 * TermScore + 3 * FuzzyScore}{8} \quad (3)$$

The composite score intentionally emphasizes term accuracy, as using correct terminology is more important than utilizing fuzzies maximally. Also, in many cases a fuzzy will contain a translation for a term, and if that happens to be different from the specified term, using the correct term lowers the fuzzy scores. Emphasizing term scores compensates for that. The composite score for the whole test suite is calculated as the average of the composite scores for individual test cases, to lessen the effect of test cases with many terms and fuzzies on the overall score.

5.2 Comparison to LLMs

LLMs have been shown to produce better translations than traditional NMT models, at least for high-resource language pairs, such as English to Spanish (Kocmi et al., 2024). LLMs can also be used for RAT by modifying the prompt used to generate the translations (Moslem et al., 2023). While LLM superiority has been shown in the field of generic MT, RAT implemented with NMT (NMT-RAT) has not been thoroughly compared to RAT implemented with LLMs (LLM-RAT). Bouthors et al. (2024) compares NMT-RAT with LLM-RAT using 1-3 fuzzy matches, and finds NMT-RAT much better, but the LLM they compare against does not represent the state of the art.

We compare our models against two recent LLMs, Gemma 3 12B and EuroLLM 9B. We choose these models as they are both competitive and relatively small. We do not compare our models against the largest available models, as our models are aimed at professional translation, where latency and the possibility to deploy models locally is important.

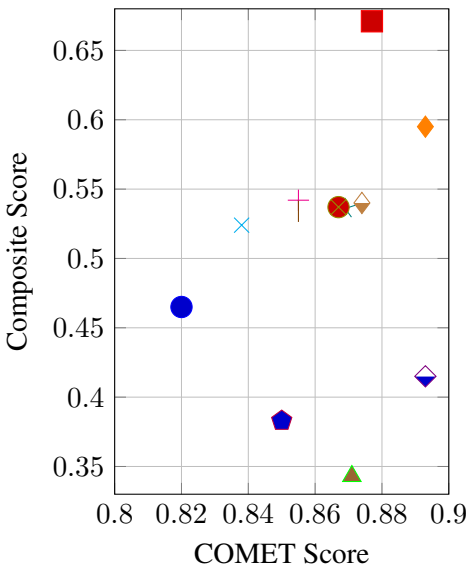
One use case that we foresee for RAT systems is interactive MT, where the MT output is influenced by translator actions. Interactive MT with RAT can for instance take the form of excluding an irrelevant match or modifying a somewhat relevant match manually during translation. This requires very fast generation of translations, as the translator should be able to see the effect of their actions almost immediately. Traditional NMT can gener-

System	Description
Baseline	Standard MT model trained without term or fuzzy annotations.
TermOnly	Model trained with 1-10 term annotations per sentence.
FuzzyOnly	Model trained with 1-3 fuzzy annotations per sentence, using both **source** and **target** similarity fuzzies.
TermAndFuzzy	Model trained with both 1-10 term and 1-3 fuzzy annotations, using both **source** and **target** similarity fuzzies.
TermAndFuzzyBig	Model trained with both 1-10 term and 1-3 fuzzy annotations, using both **source** and **target** similarity fuzzies. Transformer-big model.
ContrastEnsembleTS	Contrastive ensemble of a term and a fuzzy model (trained with both **source** and **target** similarity fuzzies). Each term and fuzzy gets own model in the ensemble.
ContrastEnsembleS	Contrastive ensemble of a term and a fuzzy model (trained with **source** similarity fuzzies). Each term and fuzzy gets own model in the ensemble.
ContrastEnsembleT	Contrastive ensemble of a term and a fuzzy model (trained with **target** similarity fuzzies). Each term and fuzzy gets own model in the ensemble.
BaselineEnsembleTS	Normal ensemble of a term and a fuzzy model (trained with both source and target similarity fuzzies). Each term and fuzzy gets own model in ensemble.
ContrastEnsembleMulti	Contrastive ensemble of a term and a fuzzy model (trained with both source and target similarity fuzzies). Ensemble contains one model for all terms, and one model for all fuzzies.
Gemma3-12B-IT	Instruction-tuned Gemma-3 LLM model with 12B parameters.
EuroLLM-9B-Instruction	Instruction-tuned EuroLLM model with 9B parameters.

Table 1: MT systems evaluated using the test suite.

ate translations fast enough, even running on desktop computers, but it is an open question whether LLMs can achieve the same. While it is not feasible currently to generate translations quickly enough locally with Gemma 3 12B and EuroLLM 9B, we use them as stand-ins for near-future LLMs that can produce translations almost immediately on desktop computers.

5.3 Results



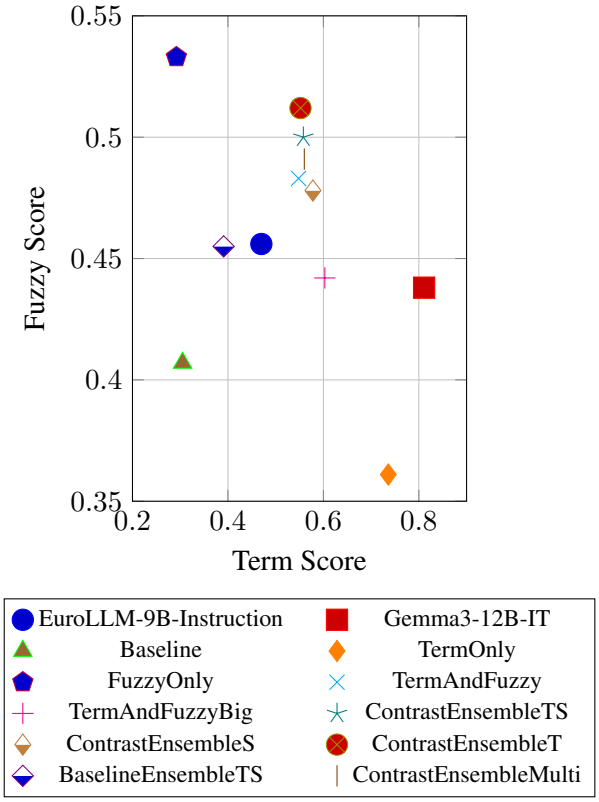


Figure 4: Term score vs Fuzzy score and COMET score vs Composite score.

We evaluated the output of 12 systems with the test suite (see models and their descriptions in Table 1). The main impression of the evaluation is that the

System	COMET ref-free	Comp. score	Term score	Fuzzy score	Fuzzy score (0 terms)
Baseline	0.871	0.343	0.305	0.407	0.520
TermOnly	**0.893**	0.595	0.736	0.361	0.481
FuzzyOnly	0.850	0.383	0.292	**0.533**	0.680
TermAndFuzzy	0.838	0.524	0.548	0.483	0.670
TermAndFuzzyBig	0.855	0.542	0.603	0.442	0.608
ContrastEnsembleTS	0.869	0.536	0.558	0.500	0.**687**
ContrastEnsembleS	0.874	0.540	0.578	0.478	0.642
ContrastEnsembleT	0.867	0.537	0.552	0.512	0.676
BaselineEnsembleTS	**0.893**	0.415	0.391	0.455	0.675
ContrastEnsembleMulti	0.855	0.534	0.560	0.491	0.659
Gemma3-12B-IT	0.877	**0.671**	**0.811**	0.438	0.639
EuroLLM-9B-Instruction	0.820	0.465	0.470	0.456	0.592

Table 2: Test suite scores for each model. The Comet model used is *wmt23-cometkiwi-da-xl*.

test suite is difficult for MT systems, with most systems performing poorly. The strongest performer by far is Gemma-12B-IT, which is also the largest evaluated model. This is a further demonstration of the edge that LLMs have over traditional NMT models, although it should also be noted that the second evaluated LLM (EuroLLM-9B-Instruction) performed worse than the NMT-RAT models. It is also noteworthy that Gemma-12B-IT excels above all in using the correct terminology in its translations. However, it does not score nearly as well in the fuzzy categories. As mentioned, high term accuracy impacts the scores of the fuzzy categories, but Gemma-12B-IT fuzzy score is lower than with the NMT-RAT systems also in cases where there are no terms in the input (see the last column in Table 2).

When comparing the NMT-RAT systems to each other, we can confirm that naive ensembling of models (BaselineEnsembleTS) results in the dilution of the impact of individual models, causing lower term and fuzzy scores. Contrastive ensembling (ContrastEnsemble models) clearly remediates this problem, although term scores remain low compared to the scores produced by the pure term model. The only transformer-big model (TermAndFuzzyBig) in the evaluation has comparable performance to the ContrastEnsemble models, which again demonstrates the effectiveness of contrastive ensembling.

The reference-free COMET scores are fairly similar across systems, with EuroLLM-9B-Instruction being the only outlier. Based on these scores, the RAT methods used do not degrade general output quality.

It is notable that the term scores are low relative to comparable previously published scores, such as those in Alam et al. (2021). This is likely due to the fact that the terms in the test suite have multiple feasible translations by design, which makes the task of applying the correct terminology more difficult.

6 Conclusion

Our experiments demonstrate that combining multiple types of information in a RAT system remains an open problem, even though LLMs show much promise also in this field. The main contributions of this paper are the introduction of contrastive ensembling and the dedicated, extensive test suite for evaluating RAT. Using the test suite we confirm that contrastive ensembling with separate term and fuzzy models provides better results than naive ensembling or single models that can process both terms and fuzzies. While contrastive ensembling does not perform as well as Gemma-12B-IT LLM, its computational requirements are much lower, which makes it suitable to more use cases, such as local low-latency RAT. We have made the training and evaluation pipeline[4] and the test suite[5] available under a permissive license.

7 Limitations

Because we rely entirely on the test suite for evaluation, our experiments are limited to one language

[4]https://github.com/Helsinki-NLP/OpusDistillery/tree/modularization

[5]https://github.com/TommiNieminen/RatTestSuite

direction. However, as our language direction is challenging due to a morphologically complex target language, we can be reasonably confident that the results also apply to less demanding language directions. The test suite has been generated semi-automatically, and while all the test items have been reviewed manually, they may not fully resemble naturally occurring data. The formulas we use to calculate the composite scores of the test suite are motivated by practical considerations, but they may place too much emphasis on certain aspects of the translations, especially correct terminology use. While reference-free MT quality metrics have been shown to work well in recent evaluations (Freitag et al., 2024), they may behave unexpectedly with unusual text types and language pairs.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the wmt shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Maxime Bouthors, Josep Maria Crego, and François Yvon. 2024. Retrieving examples from memory for retrieval augmented neural machine translation: A systematic comparison. In *NAACL-HLT*.

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chikiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are llms breaking mt metrics? results of the wmt24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2017. Search engine guided non-parametric neural machine translation. *ArXiv*, abs/1705.07267.

Cuong Hoang, Devendra Sachan, Prashant Mathur, Brian Thompson, and Marcello Federico. 2023. Improving retrieval augmented neural machine translation by controlling source and fuzzy-match interactions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 289–295, Dubrovnik, Croatia. Association for Computational Linguistics.

Hieu Hoang, Huda Khayrallah, and Marcin Junczys-Dowmunt. 2024. On-the-fly fusion of large language models and machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 520–532, Mexico City, Mexico. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Annual Meeting of the Association for Computational Linguistics*.

John Hutchins. 1998. The origins of the translator's workstation. *Machine Translation*, 13:287–307.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn and Jean Senellart. 2010. Convergence of translation memory and statistical machine translation. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 21–32, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle.

Tommi Nieminen, Jörg Tiedemann, and Sami Virpioja. 2025. Incorporating target fuzzy matches into neural fuzzy repair. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 408–418, Tallinn, Estonia. University of Tartu Library.

Franz Josef Och and Hermann Ney. 2001. Statistical multi-source translation. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Vikas Raunak, Roman Grundkiewicz, and Marcin Junczys-Dowmunt. 2024. On instruction-finetuning neural machine translation models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1155–1166, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2009. *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*, volume V, pages 237–248.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Ke Wang, Jun Xie, Yuqi Zhang, and Yu Zhao. 2023. Improving neural machine translation by multi-knowledge integration with prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5000–5010, Singapore. Association for Computational Linguistics.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. "bicleaner AI: Bicleaner goes neural". In *"Proceedings of the Thirteenth Language Resources and Evaluation Conference"*, pages "824–831", "Marseille, France". "European Language Resources Association".

A Prompt used for generating RAT output with LLMs

Both LLMs tested use the same system prompt, but the user prompt had to be customized for each model to produce clearly delineated translation output.

System prompt for both LLMs: You are a translator translating from English to Finnish.

User prompt for Gemma3-12B-IT: Translate the sentence below to Finnish using the specified terms and fuzzy matches. Use the structure of the fuzzy matches in the translation if appropriate, but do not copy parts of the fuzzy match to the translation if they are not semantically present in the source sentence. Using the specified term is more important than using the fuzzy match, so if a term and the fuzzy match conflict, always prefer the term. Output the answer in the following format, and do not output anything else:
TRANSLATION: TRANSLATION GOES HERE
Terms: *source term 1=target term 1*,*source term 2=target term 1*...
Fuzzy match 1: *target side of fuzzy match 1*
Fuzzy match 2: *target side of fuzzy match 2*...

User prompt for EuroLLM-9B-Instruction: Translate the sentence below to Finnish using the specified terms and fuzzy matches. Use the structure of the fuzzy matches in the translation if appropriate, but do not copy parts of the fuzzy match to the translation if they are not semantically present in the source sentence. Using the specified term is more important than using the fuzzy match, so if a term and the fuzzy match conflict, always prefer the term. Only output the translation.
Terms: *source term 1=target term 1*,*source term 2=target term 1*...
Fuzzy match 1: *target side of fuzzy match 1*
Fuzzy match 2: *target side of fuzzy match 2*...

A Cross-Lingual Perspective on Neural Machine Translation Difficulty

Esther Ploeger[1], **Johannes Bjerva**[1], **Jörg Tiedemann**[2], **Robert Östling**[3]
[1]Aalborg University [2]University of Helsinki [3]Stockholm University
{espl,jbjerva}@cs.aau.dk jorg.tiedemann@helsinki.fi robert@ling.su.se

Abstract

Intuitively, machine translation (MT) between closely related languages, such as Swedish and Danish, is easier than MT between more distant pairs, such as Finnish and Danish. Yet, the notions of 'closely related' languages and 'easier' translation have so far remained underspecified. Moreover, in the context of neural MT, this assumption was almost exclusively evaluated in scenarios where English was either the source or target language, leaving a broader cross-lingual view unexplored. In this work, we present a controlled study of language similarity and neural MT difficulty for 56 European translation directions. We test a range of language similarity metrics, some of which are reasonable predictors of MT difficulty. On a text-level, we reassess previously introduced indicators of MT difficulty, and find that they are not well-suited to our domain, or neural MT more generally. Ultimately, we hope that this work inspires further cross-lingual investigations of neural MT difficulty.

1 Introduction

In neural machine translation (NMT), the choice of the language pair(s) under study is often heavily influenced by data availability. But how does the choice of language pair influence the difficulty of the translation task? This question has been studied extensively for statistical MT (e.g., Koehn, 2005; Birch et al., 2008; Paul et al., 2009); these works are still frequently cited to explain linguistic disparity in MT (e.g., Rowe et al., 2025). But to date, similar studies on NMT have been scarce. A notable exception was presented by Bugliarello et al., 2020, who quantified translation difficulty for Transformer models (Vaswani et al., 2017). Yet, whereas the aforementioned studies in statistical MT focused on more than 100 translation directions, the NMT study is limited to English-centric translation scenarios: settings where English is either the source or the target language. As a result, little is known about cross-lingual NMT difficulty in a broader sense. The first goal of this work is to address this knowledge gap, by expanding the scope of analyzed language pairs (e.g., Figure 1).

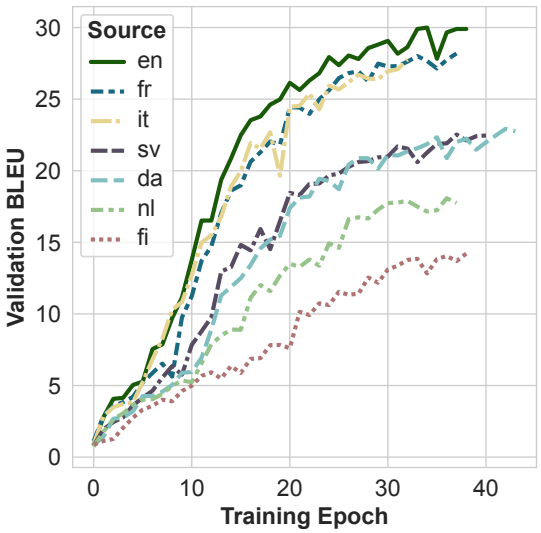


Figure 1: Training curves for MT into Portuguese; Italian (*it*) and French (*fr*) are 'easier' source languages to learn than e.g. Danish (*da*) and Swedish (*sv*).

Beyond advances in MT architectures, new approaches for quantifying language similarity have also emerged. The `lang2vec` (Littell et al., 2017) toolkit is one of the most popular typological similarity tools in natural language processing (NLP), despite major issues of data sparsity and irreproducibility (Khan et al., 2025). Typological research has resulted in newer databases, such as Grambank (Skirgård et al., 2023), which was developed with computational applications in mind (Haynie et al., 2023). This data was shown to be informative for NLP tasks, for example in the case of part-of-speech tagging (Rice et al., 2025), but there has not yet been a study on relating this new source of information to MT difficulty. The second objective of this work is therefore to systematically compare these typological databases in the context of MT.

Proceedings of the Tenth Conference on Machine Translation, pages 340–354
November 8-9, 2025

The third aim of this work is to bridge two perspectives on MT difficulty. Separately from language-level approaches to performance prediction (e.g., Birch et al., 2008; Bugliarello et al., 2020), MT difficulty has long been approached from the perspective of *text-level* difficulty (e.g., Underwood and Jongejan, 2001; Bernth and Gdaniec, 2001; O'Brien, 2004). It is not clear how these two perspectives influence each other. Are typical indicators of translation difficulty, such as sentence length, perhaps less problematic when the source and target language are more similar? In summary, this work aims to address the following three research questions:

- **RQ1:** Are there differences in NMT difficulty, beyond English-centric language pairs?
- **RQ2:** Which measures of language similarity are informative for predicting NMT performance?
- **RQ3:** How language pair-dependent are text-level translatability indicators?

Our study additionally aims to address a methodological issue in similar previous work. The inclusion of translated text in MT evaluation sets can artificially inflate results (Zhang and Toral, 2019; Graham et al., 2020), but this variable was not controlled for in relevant previous studies. To control as much as possible for variations beyond language similarity (data size, domain, topic, genre, proportion of translated text), we construct a new fully multi-parallel dataset. Because of these strict constraints and data availability, we limit our scope to 8 European languages. We train and analyze bilingual NMT models for each of the 56 resulting translation directions.

2 Related work

2.1 Machine Translation Difficulty

Estimating the difficulty of a translation task (*translatability*) from text has long been an important research direction in translation studies (Sun, 2015), for both manual and automatic translation (Bernth and Gdaniec, 2001). For MT, assessing the difficulty of translation tasks has had two prominent applications in the translation industry. The first is to distinguish samples that are suitable for MT from more difficult samples that likely require manual translation (Underwood and Jongejan, 2001). This line of research was especially popular when the performance of MT systems was much poorer than it is now. Still, more recently, Fernicola et al. (2023) showed that NMT suitability can also be predicted from source texts with reasonable accuracy. The second application has been to inform *controlled language* writing (e.g., Miyata et al., 2015). Given indicators of text characteristics that pose issues to MT, writers can choose to avoid these, to make their texts easier for MT systems to translate.

Text-level Approaches Certain text-level characteristics have been associated with MT difficulty. Examples include personal pronouns, post-modifying adjective phrases, ellipsis and very long or short sentences (Bernth and Gdaniec, 2001). Such indicators are also called *translatability indicators* (TIs). Underwood and Jongejan (2001) distinguish *general* TIs from *system-specific* ones. O'Brien (2004) writes that "it has been acknowledged that some TIs are more problematic for certain language pairs and directions than others". In the context of neural MT, however, the relevance of these TIs has not been investigated, and the influence of the language pair is also unclear. We explore a number of general TIs across multiple language pairs in Section 7.

Language-level Approaches The relationship between language similarity and MT difficulty has been a longstanding area of interest in MT research. Early work on statistical MT by Koehn (2005) and Birch et al. (2008) explored translation challenges across language pairs, using BLEU scores (Papineni et al., 2002) as a primary evaluation metric. These studies showed that translation performance tends to correlate with linguistic proximity; historically closely related languages generally yielded higher BLEU scores. Bugliarello et al. (2020) presented the first similar study on neural MT, noting that BLEU scores are only comparable on test sets in the same target language. They separated source- and target-language difficulty explicitly, through measuring cross-mutual information, with evaluations on 40 English-centric language pairs.

2.2 Language Similarity in MT

How to best measure language similarity to inform NLP research is an open question. Blaschke et al. (2025) conducted a large-scale study on distance measures for cross-lingual transfer in three NLP tasks: dependency parsing, part-of-speech tagging and topic classification. They found that the definition of linguistic similarity is an important factor

for cross-lingual transfer success, and the most effective similarity measure is dependent on the downstream task. Since MT was not investigated, it is not clear which kind of similarity measure is most informative for this task. Within MT research, language similarity has mostly been of interest for improving transfer learning performance (e.g., Lin et al., 2019; Oncevay et al., 2020; Fekete et al., 2025). Factors beyond language similarity can play a major role in such investigations, such as data availability and domain similarity (Khiu et al., 2024). By contrast, we are interested specifically in the difficulty of the translation task by itself, all other factors being as equal as possible.

3 Constructing a Multi-Parallel Corpus

Our aim is to investigate the difficulty of translation tasks, while controlling for factors other than language similarity. We first establish three criteria that should be met for comparable cross-lingual MT studies (Section 3.1). We then describe how these were accounted for in the creation of our dataset (Section 3.2), and how these restrictions impact the diversity of our language selection (Section 3.3).

3.1 Criteria

If MT models for different language pairs are trained on different datasets, then differences in results could be attributed to that instead of linguistic divergences. So, to rule out the effects of dataset size and domain differences, the dataset should be fully multi-parallel across all included languages. Secondly, within the multi-parallel corpus, the distribution of 'original text' should be equal across languages. A well-described problem in the evaluation of MT systems is that the presence of translated data in the evaluation set can inflate performance assessments (Zhang and Toral, 2019; Graham et al., 2020).[1] Ideally, test sets should contain text originally written in a language, to not exhibit 'translation artefacts'.[2] However, since such a dataset should also be completely multi-parallel, this is not possible for more than one language. We argue that a dataset should therefore ensure equal *proportions* of translated text in the multi-parallel test set. In addition to consistent proportions across languages, it should be the same across training and testing sets, to avoid training on one text type and evaluating another. Lastly, to provide a cross-lingual perspective beyond English, the dataset should contain translation pairs that do not include English as either source or target language. In summary, we establish three criteria:

1. Full multi-parallelism
2. Equal translated text distribution
3. Beyond English-centric MT

To the best of our knowledge, a ready-to-use dataset that satisfies these three criteria does not yet exist.

3.2 Dataset Creation

Multiple popular multi-parallel datasets with broad language coverage exist (e.g., Parallel Bible Corpus, Mayer and Cysouw, 2014; OpenSubtitles, Lison and Tiedemann, 2016). Yet, these datasets do not contain information on the original languages of the data. The CoStEP corpus (Graën et al., 2014) includes speaker turns from the European Parliament, cleaned and aligned across languages, with original-language annotations. In Figure 2, we illustrate the steps for creating a dataset that satisfies the criteria from Section 3.1. Starting from the CoStEP corpus (1) we first extract all bilingually aligned speaker turns (2). Since speaker turns vary in length (some are very long while others are very short), we split the turns into sentences (3) using the `sentence-splitter` toolkit.[3] Then, we re-align the bilingual parallel sentences (4) using `hunalign` (Varga et al., 2008). Note that we refrain from using embedding-based models like VecAlign (Thompson and Koehn, 2019), to avoid potential cross-lingual biases (English-centric pairs having higher-quality alignments, because English is very well-represented in pre-trained models). Next, given these aligned sentences per original language, we assess for which languages we have enough multi-parallel data for MT training and evaluation (5). We find that there are eight languages that have both high coverage and meta-information available in the typological database Grambank (Skirgård et al., 2023), which we use because of its suitability for computational applications (Haynie et al., 2023): Danish, Dutch, English, Finnish, French, Italian, Swedish and Portuguese. We select all samples for which we have translations in all languages and make a multi-parallel dataset with equal original-language proportions (6). Lastly, we divide the text

[1] In the MT literature, this is commonly described as the 'translationese effect', while this term is not uncontroversial (Jimenez-Crespo, 2023).

[2] We empirically verify this for our dataset in Appendix A.

[3] https://github.com/mediacloud/sentence-splitter

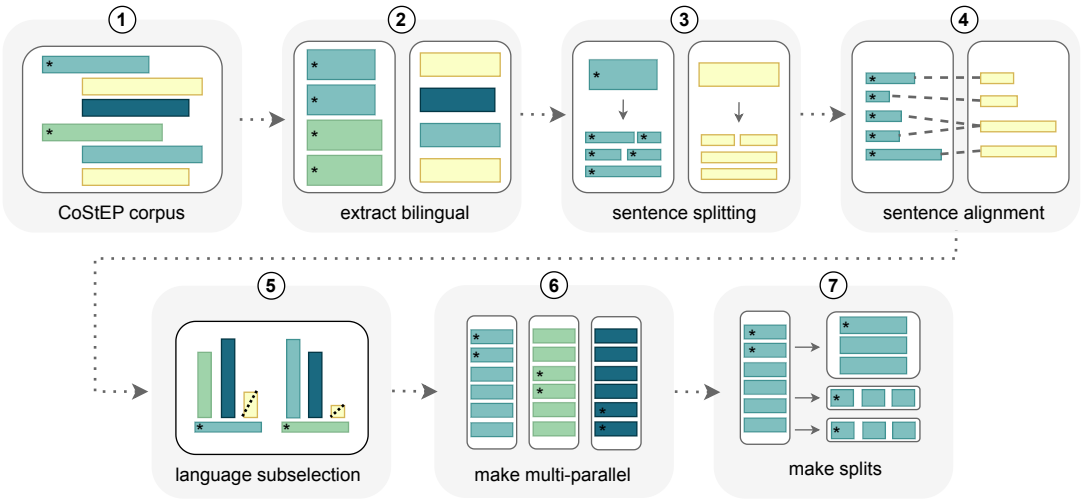


Figure 2: Schematic overview of the dataset creation process, starting from the CoStEP corpus, where we ensure an equal original text distribution (indicated schematically with an asterisk) among the included languages.

into splits for MT training, validation and testing such that the original text proportion remains equal among languages (7). We reserve 800 samples for validation (100 original per language) and 1,600 for testing (200 original per language). We use the remaining lines (77,941 per language) for training.

The number of space-separated tokens per language (as obtained by running the `wc -w` command) is in Table 1. Here, it stands out that Finnish contains a lower number of 'words' than the other languages, as it exhibits more morphologically complexity than the others.

Language	Train	Valid	Test
Danish (da)	1.9M	20.0K	39.3K
Dutch (nl)	2.1M	21.8K	43.0K
English (en)	2.1M	22.0K	42.6K
Finnish (fi)	1.4M	14.8K	28.9K
French (fr)	2.2M	23.0K	44.6K
Italian (it)	2.0M	21.2K	40.9K
Swedish (sv)	1.9M	22.2K	43.0K
Portuguese (pt)	2.1M	19.8K	38.8K

Table 1: Number of 'words' per language in the multi-parallel dataset.

Our strict filtering criteria result in a controlled, but small dataset which is not necessarily representative of state-of-the-art high-resource NMT more generally. These strict controls are necessary for our study, and we retrieve reliable results within our set-up, but we cannot make claims regarding the broad generalizability of these results.

3.3 Typological Diversity

Our language selection contains languages from three genera (Germanic: Danish, Swedish, Dutch, English; Romance: Italian, French, Portuguese; and Finnic: Finnish). To gain a more fine-grained image beyond genealogical similarities, we look into the typological similarity of these languages using the Grambank database (Skirgård et al., 2023).[4] Figure 3 shows a PCA plot of the Grambank feature vectors, showing our language selection is not representative of typological diversity generally. While this limited language diversity, also in terms of writing systems, means we cannot make any claims about what makes NMT difficult *in general*, it is appropriate for the objective of providing cross-lingual (but not universal) insights.

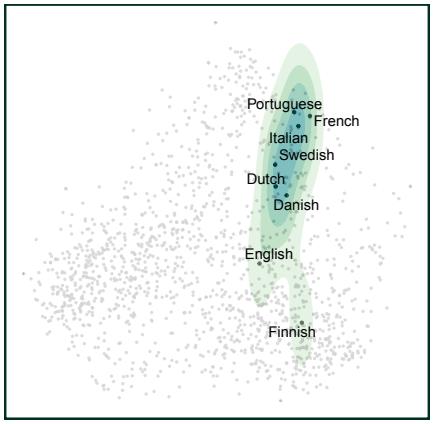


Figure 3: PCA plot from Grambank features, illustrating the (limited) typological diversity in this study.

[4]Typological diversity using *typdiv* (Ploeger et al., 2025, default settings): MPD: 0.51, FVO: 0.79, FVI: 0.77, H: 0.42.

T (→) / S (↓)	da		nl		en		fi		fr		it		sv		pt	
	chrF	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*	*chrF*	*BS*
da	–	–	69.9	**0.81**	74.3	0.85	67.3	0.78	72.2	0.82	70.1	0.80	**75.2**	**0.85**	71.4	0.82
nl	69.8	0.82	–	–	**69.8**	**0.82**	60.9	**0.72**	66.9	0.77	**62.8**	**0.76**	**67.4**	0.80	67.5	0.80
en	74.6	0.84	**71.3**	**0.81**	–	–	**70.0**	**0.80**	75.4	**0.84**	72.6	**0.83**	74.0	0.84	**75.6**	**0.85**
fi	**68.1**	**0.79**	**62.8**	**0.73**	**69.8**	0.83	–	–	**58.2**	**0.73**	68.1	0.77	68.0	0.80	**66.8**	**0.78**
fr	70.9	0.82	67.5	0.79	74.0	0.85	**60.5**	0.73	–	–	73.9	**0.83**	70.3	0.81	74.7	0.84
it	68.8	0.80	67.7	0.77	73.3	0.84	65.3	0.76	72.0	0.82	–	–	67.8	**0.79**	73.9	0.83
sv	**75.1**	**0.85**	70.8	0.80	75.0	**0.86**	68.5	0.79	72.3	0.82	71.2	0.80	–	–	71.3	0.82
pt	71.1	0.82	67.8	0.79	**76.3**	**0.86**	66.9	0.77	**75.8**	**0.84**	**74.8**	**0.83**	72.3	0.82	–	–

Table 2: chrF2 and BERTScore for each target language (columns), per source language (rows). **Highest** and **lowest** scores per metric are highlighted for each target language.

4 Machine Translation Models

For each language separately, we train a Unigram (Kudo, 2018) subword segmenter on the training corpus, with vocabulary size 8,000, to tokenize the text. We then train a separate bilingual MT model for each of the 56 language pairs in our dataset. We choose to train small models from scratch, rather than leveraging pre-trained NMT models or LLMs, to avoid cross-lingually unfair pre-training distributions. The goal is not to create optimally functioning systems, which would require intensive parameter tuning, but rather to compare systems that learn to translate between different language pairs under otherwise identical circumstances. We use the Fairseq toolkit (Ott et al., 2019) to train MT models with a standard Transformer (Vaswani et al., 2017) architecture, consisting of 6 encoder and 6 decoder layers. We use a learning rate of 0.0001 and a dropout rate of 0.2. We use a cross entropy with a label smoothing of 0.2. and a maximum of 4,000 tokens per training batch. As a best checkpoint metric we use cross-entropy, with patience 5. No model took more than 50 epochs to converge.

5 Cross-lingual Translatability (RQ1)

5.1 Measuring Translation Difficulty

We approach translation difficulty from two perspectives: MT performance estimates, and the computational resources required for training.

Translation Accuracy Inspired by works from human translatability, e.g., Vanroy et al. (2019) and Hale and Campbell (2002), as well as MT approaches (Koehn, 2005; Birch et al., 2008), we deem a translation task difficult if it triggers errors. The intuition is that language pairs that are more difficult to translate lead to lower translation accuracy. As a widely spread measure of translation quality, we report chrF2 (Popović, 2015) scores (Table 2), which were previously shown to be robust against varying degrees of morphological complexity (Popović, 2016). This is a surface-level metric, based on a reference translation. We use SacreBLEU (Post, 2018) to calculate this.[5] One shortcoming of these measures, is that wording differences in the human reference translations can influence the results. This is why we also ran an embedding-based evaluation. We do not use reference-free metrics such as COMET (Rei et al., 2020), since the unequal training data in language embeddings may introduce cross-lingual unfairness in evaluation. Instead, we compare (monolingually) the MT hypothesis against the reference using BERTScore (Zhang et al., 2019), which renders it comparable per target language.

Computational Resources Inspired by works on human translation and post-editing that measured translation difficulty through cognitive and temporal effort (e.g., Campbell, 1999; Beinborn, 2010), we deem a translation direction more difficult if it requires more resources. We examine "machine effort" through a proxy: training dynamics. Beyond the overall performance, we record the loss and BLEU on the validation set per epoch as measures of effort, and compare this across source languages for MT into the same target language.

5.2 Difficulty as Translation Accuracy

Overall Performance We list the chrF2 and BERTScore per target language in Table 2. We exclusively compare the source languages per target language, since direct comparison across different target languages' test sets may be unfair (Bugliarello et al., 2020). Higher scores indicate higher translation accuracy, and thereby suggest lower MT difficulty. While the absolute scores are

[5] signature: `nrefs:1|case:mixed|eff:yes|nc:2|nw:0|space:no|version:2.5.0"`

low compared to the state-of-the-art, we observe that, indeed, there are language-level translatability differences. In other words, it is not the case that the same source language is the easiest for all target languages. For example, the easiest source language for MT into Danish is Swedish, while the easiest for Portuguese is French. Finnish, as the only language outside of the Indo-European language family in this study, also stands out in terms of MT performance. For most target languages (Danish, Dutch, English, French, Portuguese), it is among the most difficult. Moreover, we see that Dutch is a difficult source language for many target languages. These results indicate that, also in scenarios beyond English-centric MT, translation difficulty varies translation per direction. We observe some intuitive patterns, such as Danish↔Swedish being a relatively easy translation direction. In Section 6, we assess the connection with language similarity more systematically.

5.3 Difficulty as Effort

Training Epochs for Validation BLEU Analogous to "the extent to which cognitive resources are consumed by a translation task for a translator to meet objective and subjective performance criteria." (Sun, 2015), we assess how many computational resources ("machine effort") are required for an objective performance criterion. Specifically, we assess how many training epochs are needed to reach 15 BLEU on the validation set. While this threshold is somewhat arbitrary, we note that the full training progressions are shown in Appendix C. While this metric cannot be compared across languages, due to its reliance on word boundaries, it provides an intuitive and easy to interpret measure of difficulty. Here, a lower number (of epochs) indicates that fewer computational resources are required, signaling that the translation direction is easier. Table 3 shows the results. Firstly, we observe that this analysis reveals different nuances than when approaching difficulty as accuracy. Namely, English is an 'easy source language' for more languages here. As such, generalizing English-centric translation results may lead to overestimating NMT performance. This is especially noteworthy, as most previous studies (and MT in general) center research claims around this language. An explanation for this, relative to our dataset, could be that the European Parliament sometimes uses relay translations, i.e. manual translations are produced through English as a pivot language. Unfortunately, it is not possible to control for this variable.

Furthermore, it is interesting to analyze the translation directions which never reach 15 validation BLEU, indicated by the ∅ symbol. Especially Finnish stands out here, as it only reaches the threshold from English. Additionally, none of the Romance target languages reach 15 BLEU when Finnish is the source language. These are patterns that follow intuitive language similarity ideas, which are further explored in Section 6.

Loss Slopes While intuitive, the 15 BLEU threshold has shortcomings. For one, it is only comparable per target language, due to its reliance on word boundaries. Here, we analyze a more cross-lingually comparable metric: the loss slope between two consistent, pre-defined points in the training process (5^{th} epoch vs. 25^{th} epoch). Starting from epoch 5 gives the decoders a chance to learn a very basic language model, ahead of actually learning to translate. In Appendix D, the full slopes are visualized. Higher numbers indicate steeper slopes, which indicates faster learning, and

S (↓)	da	nl	en	fi	fr	it	sv	pt
da	–	19	12	∅	15	29	**12**	19
nl	19	–	16	∅	32	∅	29	27
en	**12**	**16**	–	**32**	**11**	16	14	**12**
fi	31	∅	18	–	∅	∅	34	∅
fr	19	25	12	∅	–	15	20	14
it	30	∅	12	∅	14	–	35	13
sv	13	22	12	∅	17	26	–	18
pt	18	29	**9**	∅	12	**14**	24	–

Table 3: Number of training epochs needed to reach at least 15 BLEU on the validation set per translation direction, lowest number per target language **highlighted**.

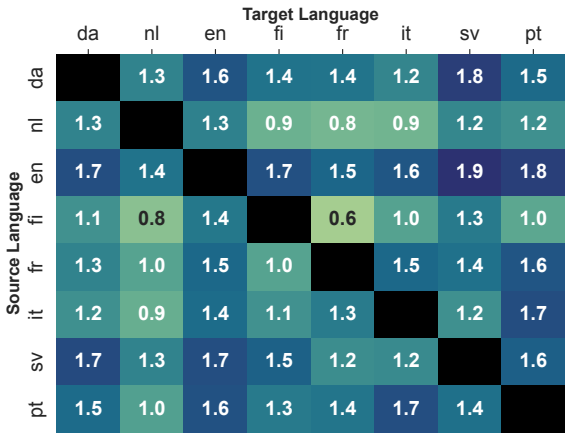
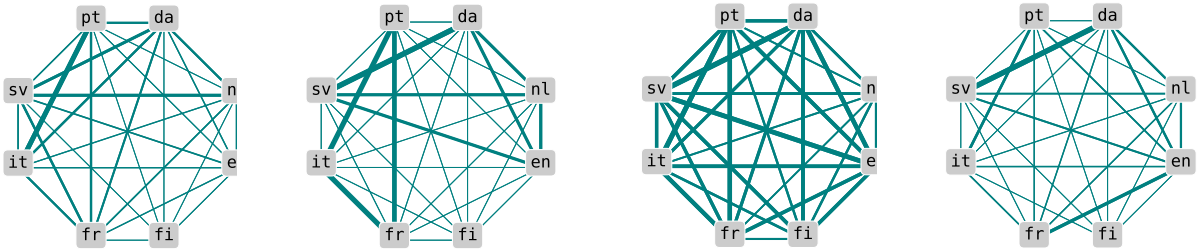


Figure 4: Δ losses (5^{th} vs. 25^{th} epoch) in MT training.



(a) Grambank similarity (b) Genealogical similarity (c) Word order similarity (d) Subword overlap

Figure 5: Visual comparison of language-based (a, b) and data-based (c, d) language similarity, where line thickness illustrates similarity. For example, subword overlap between Danish and Swedish is particularly high (d).

thus suggests lower translation difficulty. In Figure 4, the steepest slopes (darker colors) are mostly found for English. This is in line with our findings from using the validation BLEU threshold, where English also was commonly 'easy'. Additionally, it stands out that Dutch is relatively difficult. One possible explanation for this is found in Figure 5c, where Dutch generally has the most dissimilar word order to the other languages. We verify this relationship in Section 6.

All in all, we conclude that there are indeed language pair-level differences in NMT translation difficulty. This indicates that the vague notion of 'easier translation' can be operationalized more systematically. Results on translation difficulty, both from the angles of accuracy and effort, show intuitive patterns with regard to language similarity patterns, which are further examined in the next section.

6 Language Similarity Analysis (RQ2)

6.1 Measuring Language Similarity

Any effort to reduce languages and their similarities to single floating point numbers risks being simplistic. Yet, to enable systematic comparison, we need to compare the different aspects of similarity on the same scale. We compare two categories of metrics: those derived on a language-level, and dataset-specific measures.

Language-Level Metrics A first approach is to determine language similarity based on expert annotations, such as phylogenies and typological features. The popular `lang2vec` toolkit provides six categories of language distances (geographical, genealogical, syntactic, phonological, featural and inventory-based), based on the URIEL database (Littell et al., 2017). Given these distances, d, the resulting language similarity is then defined as $1 - d$. We compare this with a Grambank-based (Skirgård et al., 2023) measure, as proposed in Ploeger et al. (2025): the Euclidean distance between Grambank's morphosyntactic feature vectors, accounting for missing values. We again subtract this from 1, to compute a similarity score.

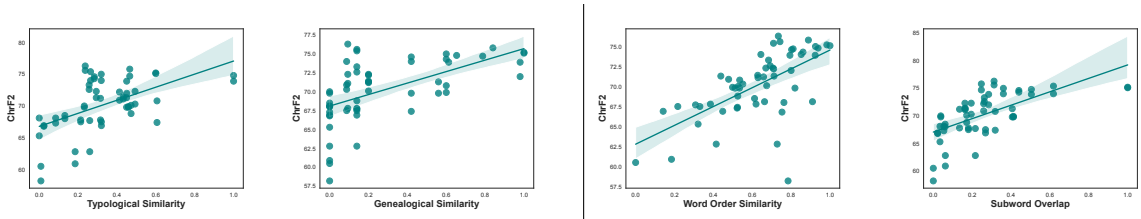
Text-Driven Metrics As a more data-specific measure of *syntactic* similarity, we use the relative amount of word reordering between two languages, calculated over the test set using Eflomal (Östling and Tiedemann, 2016). Since reordering scores are directional, we take the average over both directions to retrieve a single similarity score, in line with the language-level metrics. To go from reordering to word order similarity, we again subtract it from 1. As a *lexical* measure of data-driven similarity, we take the proportion of overlapping subwords from the tokenized texts, relative to the sum of the number of subwords of both languages in the test set. Finally, we apply MinMax scaling.

6.2 Results and Analysis

Table 4 shows the Pearson correlation coefficients for each of the language similarity metrics with

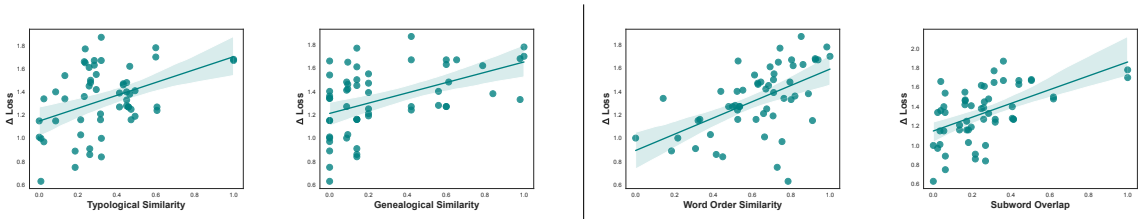
Similarity Measure	chrF2	Δ Loss
12v (geographic)	0.41*	0.26
12v (genetic)	**0.56***	**0.47***
12v (syntactic)	0.46*	0.43*
12v (featural)	-0.08	0.10
12v (inventory)	-0.18	-0.22
12v (phonological)	-0.04	-0.04
Grambank	**0.52***	**0.41***
Word reordering	**0.63***	**0.54***
Subword overlap	**0.63***	0.53*

Table 4: Pearson correlation coefficients between similarity metric and performance. Statistically significant values ($p < 0.005$) are indicated with an asterisk.



(a) Typological similarity (gb) ($r \approx \mathbf{0.52}, p < 0.005$)
(b) Genealogical similarity ($r \approx \mathbf{0.56}, p < 0.005$)
(c) Word order similarity ($r \approx \mathbf{0.63}, p < 0.005$)
(d) Subword overlap ($r \approx \mathbf{0.63}, p < 0.005$)

Figure 6: Correlation between linguistic similarity measures and chrF2 scores.



(a) Typological similarity (gb) ($r \approx \mathbf{0.41}, p < 0.005$)
(b) Genealogical similarity ($r \approx \mathbf{0.47}, p < 0.005$)
(c) Word order similarity ($r \approx \mathbf{0.54}, p < 0.005$)
(d) Subword overlap ($r \approx \mathbf{0.53}, p < 0.005$)

Figure 7: Correlation between linguistic similarity measures and Δ training loss.

chrF2 and Δ loss. From the table, it follows that morphosyntactic distance based on Grambank is a better predictor of MT difficulty than `lang2vec`'s syntactic distance. Still, `lang2vec`'s genetic distance and the data-driven measures yield higher correlations. We now examine these four metrics with the strongest correlations (Table 4, bold font) in more detail. These language similarity metrics are correlated positively with chrF2 and the Δ loss. That is, the more similar the languages in the translation direction, the 'easier' it is to translate between them. All correlations are statistically significant, with $p < 0.005$. Interestingly, the correlation coefficients of the data-driven metrics (c, d) are higher than those of the language-level metrics (a, b). This indicates that tailoring language similarity measures to the dataset under study may be beneficial for retrieving accurate difficulty predictions.

To gain an insight into why these correlations differ, we qualitatively assess the language similarities. Figure 5 illustrates the pairwise distances per metric. Absolute numbers are given in Appendix B. In the Grambank-based measure (a), the link between Italian and Portuguese is especially pronounced, while with the genealogical metric (b), we for example see a strong similarity between Portuguese, Italian and French. In terms of the data-driven measures, what stands out in (c) is that Dutch has relatively low word order similarity scores with all other languages, and that the subword overlap (d) between Danish and Swedish is especially prominent. This influences the correlation coefficients. As visualized in Figure 6 (for difficulty as accuracy; chrF2) and Figure 7 (for difficulty as effort; Δ loss), outliers –Portuguese and Italian, Danish and Swedish– influence the correlation coefficients, notably in Figure 6/7a and 6/7d respectively.

7 Translatability Indicators (RQ3)

As mentioned in Section 2, approaching translation difficulty from text on a data-level has been an important research direction in the past, mostly in statistical MT (Bernth and Gdaniec, 2001). Which textual characteristics make MT difficult? We revisit source-text indicators of translation difficulty that were proposed in previous work (Underwood and Jongejan, 2001), and contribute an investigation of these indicators in the context of neural MT.

7.1 Identifying TIs

We reassess 'general indicators' of machine translatability from Underwood and Jongejan (2001), listed in Table 5. For this case study, we take a closer look at translation with English as a source, since these TIs were formulated for that language specifically. We obtain the TIs through POS tags and dependency relations, retrieved through Trankit (Nguyen et al., 2021), with the default XLM-RoBERTa as the underlying model. We follow the heuristics defined in Underwood and Jongejan

Translatability Indicator	#	da	nl	fi	fr	it	sv	pt
No verb	11	+13.85	+12.71	+14.79	+19.30	+12.67	+13.13	+13.11
No finite verb	13	+12.92	+10.04	+12.38	+14.84	+11.43	+12.31	+13.28
Long (> 25 words)	713	**-0.21**	**-0.14**	**-0.36**	+0.52	+0.43	+0.01	+0.30
Short (< 3 words)	4	+19.7	+24.98	+27.71	+32.75	+0.68	+18.15	+25.45
$\geq$ 1 nominal compound	520	**-0.09**	**-0.13**	+0.06	+0.55	+0.56	+0.30	+0.38
Multiple coordination	759	**-0.18**	**-0.30**	+0.14	**-0.20**	+0.48	+0.43	**-0.35**

Table 5: Delta between the average chrF2 score for the TI lines, and the average non-TI chrF2. Negative values (bold font) imply increased translation difficulty. # Indicates how often the TI appears in our English test set.

(2001) for detecting the translatability indicators. Our implementation is as follows:

Missing Verbs A line does not have a verb if it contains no token with an `AUX` or `VERB` tag. An example from our dataset is: *"All well and good."* A line does not contain a finite verb if there is no verb with `VerbForm=Fin` present. An example: *"But what about all the other protective considerations listed in Article 13 of the Treaty?"*.

Sequence Length Long and short sentences are determined through the number of words: long sentences contain more than 25 words, while short sentences contain less than 3. While long sequences are ubiquitous in our dataset, an example of a (much rarer) short line is: *"No one!"*.

$\geq$1 Nominal Compound A sentence contains a nominal compound if it contains a `NOUN` that has the dependency relation of `compound`, for example: *"But we also need better regulation and principles for future EU legislation when it comes to motor vehicles."*.

Multiple Coordination Lastly, we detect multiple coordination if the sentence contains more than one `SCONJ` and/or `CCONJ`. An example: (e.g. *"That is why, of course, donor cards should be voluntary, and the same applies to the European donor card, which we intend to introduce in our action plan."*).

We first identify which samples in the English test set contain these markers. Then, we calculate the chrF2 score per sample in the test set, and compare the average score of the TIs with the average score of the non-TIs. We expect that if a certain group of sequences (TIs) is more difficult to translate, it yields a lower chrF2 score than the average non-TI chrF2 score per sample. Table 5 shows the difference between the average TI score, minus the average non-TI score. If a table cell contains a negative number (bold), this indicates that the translation of the TI samples obtained a lower chrF2 score than the others, indicating potential translatability issues.

7.2 Results and Analysis

We find no consistent patterns indicating that the defined TIs are more difficult for MT than non-TIs. For some TIs, there is too little evidence to base any robust conclusion on (e.g. only 4 lines contain $<$ 3 words in our dataset) and verb-less lines are rare. This may differ per studied domain. Secondly, it could be that common indicators of translation difficulties do not apply to our systems, because the training data is domain-specific. For example, nominal compounds occur relatively frequently in our dataset, as a result of the domain (e.g. *"member states"*, *"employment plans"*, *"terrorist list"*, *"labour taxes"*). If a model is trained on many of these, this may result in better capabilities to deal with such features. Another possible reason is that the paradigm shift to neural MT has weakened the impact of formerly informative TIs. For example, Transformers' (Vaswani et al., 2017) cross-attention implies that long-range dependencies have become less problematic.

Beyond general challenge sets, future work could be dedicated to defining TIs specifically for neural MT (cf. Bisazza et al., 2021): are there common translatability issues for state-of-the-art architectures, and consistently across datasets? Special care could be taken to make these more cross-lingually comparable: for example, taking the number of words as length indicator (Underwood and Jongejan, 2001) is dependent on the morphological complexity of a language.

8 Conclusions

In this work, we aimed to operationalize language similarity and translation difficulty in the context of neural MT. We control for confounding factors from previous work, and use the resulting dataset to answer three research questions about how language similarity affects translatability. In summary, we find the following. Firstly, there are language pair-level differences in NMT difficulty in our experiments, beyond English-centric scenarios (RQ1). Moreover, NMT difficulty can be predicted from language pair similarity with reasonable success, with syntactic and genetic measures of similarity. Text-driven metrics, tailored to the dataset, are even more informative (RQ2). Lastly, we found that text-level indicators of MT difficulty from previous work were not suitable for our dataset or evaluation set-up (RQ3).

Our models achieve limited performance (approx. 60-70 chrF2). This is because we train models on relatively small datasets. This performance is far below the state-of-the-art of these European language pairs. Still, it provides interesting insights. We show that some language pairs are reliably and consistently more difficult (e.g. much lower chrF2 scores) than others, under the same, controlled circumstances. These results are stable, as demonstrated by the steady train loss decrease in Appendix D. Furthermore, the performance differences between translation directions are substantial and predictable ($p<0.005$, Fig. 6, 7), showing that even in a limited setting, consistent results emerge. Although these strict controls are necessary for our controlled study and we retrieve reliable results in our case study, we cannot make claims regarding the broad generalizability of these results. Generalizability is a challenging topic in MT more generally, as it is unclear whether scaling approaches will always solve previously encountered issues.

These findings open up various possible applications and future research directions. For example, future work could investigate to what extent MT difficulty is influenced by tokenization strategies. In our experiments, we kept the tokenizer and vocabulary size consistent across languages, but variations could yield different results. In downstream scenarios, systematic notions of language similarity could be used to select pivot languages, especially in data-scarce scenarios. On a text-level, we showed that identifying new TIs, relevant to neural MT, may be an interesting research direction.

Limitations

Several limitations of this work should be noted. Firstly, due to the strict constraints (full multi-parallelism, equal distribution of original text), the typological diversity of the languages in our study is limited. Our findings may not generalize to other translation directions. For the same reason, we only trained and evaluated in one domain; this cannot be assumed to generalize directly either. Note that previous studies on translation difficulty (Koehn, 2005; Birch et al., 2008; Bugliarello et al., 2020) were also limited to this domain. Despite these strict constraints, 100 percent clean data is not guaranteed. For one, automatic alignments may still include noise. Also, the European Parliament sometimes includes relay translations, meaning that certain translated texts may have been translated 'through English', which could impact the results. Yet, as this information is not available in CoStEP or elsewhere, we cannot control for this. While the Transformer architecture is representative for neural MT, our study only uses this one neural architecture. For broader insights into neural MT as a whole, more approaches could be investigated. Furthermore, the language similarity measures that we evaluate are not exhaustive; for example, semantic language similarity (for example in the form of colexification) was not taken into account. Lastly, no human evaluation was included. This choice was made to ensure cross-lingually consistent comparisons, but it could still have yielded important novel insights, for example comparing human translatability and NMT features (cf. Lim et al., 2024).

Acknowledgements

We thank the AAU-NLP group, in particular Mike Zhang, for proofreading earlier versions of this article. EP and JB are funded by the Carlsberg Foundation, under the *Semper Ardens: Accelerate* Programme (project nr. CF21-0454). EP was further supported by a travel grant from the Otto Mønsteds Fond.

References

Lisa Beinborn. 2010. Post-editing of statistical machine translation: A crosslinguistic analysis of the temporal, technical and cognitive effort. Master's thesis, Saarland University.

Arendse Bernth and Claudia Gdaniec. 2001. Mtranslatability. *Machine translation*, 16:175–218.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.

Verena Blaschke, Masha Fedzechkina, and Maartje Ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8653–8684, Vienna, Austria. Association for Computational Linguistics.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649, Online. Association for Computational Linguistics.

Stuart Campbell. 1999. A cognitive approach to source text difficulty in translation. *Target. International Journal of Translation Studies*, 11(1):33–63.

Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptiste, Raj Dabre, Johannes Bjerva, and Heather Lent. 2025. Limited-resource adapters are regularizers, not linguists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 222–237, Vienna, Austria. Association for Computational Linguistics.

Francesco Fernicola, Silvia Bernardini, Federico Garcea, Adriano Ferraresi, and Alberto Barrón-Cedeño. 2023. Return to the source: Assessing machine translation suitability. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 79–89, Tampere, Finland. European Association for Machine Translation.

Johannes Graën, Dolores Batinić, and Martin Volk. 2014. Cleaning the europarl corpus for linguistic applications. In *Konvens 2014*. Stiftung Universität Hildesheim.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Sandra Hale and Stuart Campbell. 2002. The interaction between text difficulty and translation accuracy. *Babel*, 48(1):14–33.

Hannah J. Haynie, Damián Blasi, Hedvig Skirgård, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank's typological advances support computational research on diverse languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 147–149, Dubrovnik, Croatia. Association for Computational Linguistics.

Miguel A. Jimenez-Crespo. 2023. "translationese" (and "post-editese"?) no more: on importing fuzzy conceptual tools from translation studies in MT research. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 261–268, Tampere, Finland. European Association for Machine Translation.

Aditya Khan, Mason Shipton, David Anugraha, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2025. URIEL+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6952, Abu Dhabi, UAE. Association for Computational Linguistics.

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from

movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of Machine Translation Summit XV: Papers*, Miami, USA.

Minh Van Nguyen, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Sharon O'Brien. 2004. Machine translatability and post-editing effort: How do they relate. In *Proceedings of Translating and the Computer 26*, London, UK. Aslib.

Arturo Oncevay, Barry Haddow, and Alexandra Birch. 2020. Bridging linguistic typology and multilingual machine translation with multi-view language representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2391–2406, Online. Association for Computational Linguistics.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*, 106:125–146.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the importance of pivot language selection for statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 221–224, Boulder, Colorado. Association for Computational Linguistics.

Esther Ploeger, Wessel Poelman, Andreas Holck Høeg-Petersen, Anders Schlichtkrull, Miryam de Lhoneux, and Johannes Bjerva. 2025. A principled framework for evaluating on typologically diverse languages. *To appear in Computational Linguistics*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2016. chrF deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504, Berlin, Germany. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

Jacqueline Rowe, Edward Gow-Smith, and Mark Hepple. 2025. Limitations of religious data and the importance of the target domain: Towards machine translation for Guinea-Bissau creole. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 183–200, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Hedvig Skirgård, Hannah J Haynie, Harald Hammarström, Damián E Blasi, Jeremy Collins, Jay

Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, and 1 others. 2023. Grambank v1.0.

Sanjun Sun. 2015. Measuring translation difficulty: Theoretical and methodological considerations. *Across languages and cultures*, 16(1):29–54.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Nancy Underwood and Bart Jongejan. 2001. Translatability checker: a tool to help decide whether to use MT. In *Proceedings of Machine Translation Summit VIII*, Santiago de Compostela, Spain.

Bram Vanroy, Orphée De Clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *Perspectives*, 27(6):924–941.

Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2008. Parallel corpora for medium density languages. In *Recent advances in natural language processing IV: selected papers from RANLP 2005*, pages 247–258. John Benjamins Publishing Company.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Translation Performance for Source-original and Target-original Test Lines

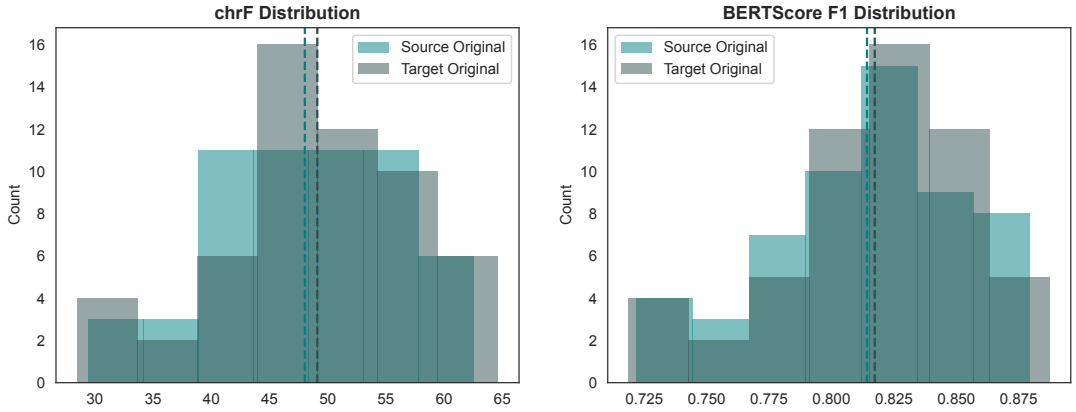


Figure 8: For each of the 56 translation directions in our study, we compute the average chrF2 and BERTScore (F1), for only those test samples that were originally spoken in the source language, with those that were originally spoken in the target language. Comparing these distributions, we observe that, in line with previous work (Zhang and Toral, 2019), scores for the target-original portion are on average higher, implying that "translationese" in the source text can inflate MT performance, albeit slightly. For this reason, we control for the proportion of translated text in our experiments.

B Pairwise Language Similarity Estimates

	Typological Similarity (Grambank)									**Genealogical Similarity (lang2vec)**							
	da	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**		**da**	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**
da	–	0.46	0.28	0.08	0.41	0.47	0.60	0.43	da	–	0.60	0.60	0.00	0.20	0.20	1.00	0.20
nl	0.46	–	0.45	0.18	0.32	0.26	0.61	0.21	nl	0.42	–	0.56	0.00	0.14	0.14	0.42	0.14
en	0.28	0.45	–	0.23	0.26	0.26	0.32	0.24	en	0.42	0.56	–	0.00	0.14	0.14	0.42	0.14
fi	0.08	0.18	0.23	–	0.01	0.00	0.13	0.02	fi	0.00	0.00	0.00	–	0.00	0.00	0.00	0.00
fr	0.41	0.32	0.26	0.01	–	0.45	0.49	0.47	fr	0.09	0.09	0.09	0.00	–	0.61	0.09	0.79
it	0.47	0.26	0.26	0.00	0.45	–	0.32	1.00	it	0.14	0.14	0.14	0.00	0.98	–	0.14	0.98
sv	0.60	0.61	0.32	0.13	0.49	0.32	–	0.29	sv	1.00	0.60	0.60	0.00	0.20	0.20	–	0.20
pt	0.43	0.21	0.24	0.02	0.47	1.00	0.29	–	pt	0.09	0.09	0.09	0.00	0.84	0.65	0.09	–

	Word Order Similarity									**Subword Overlap**							
	da	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**		**da**	**nl**	**en**	**fi**	**fr**	**it**	**sv**	**pt**
da	–	0.50	0.86	0.52	0.71	0.67	0.98	0.63	da	–	0.41	0.44	0.05	0.26	0.17	1.0	0.16
nl	0.53	–	0.51	0.18	0.45	0.41	0.52	0.33	nl	0.41	–	0.41	0.06	0.27	0.22	0.32	0.18
en	0.80	0.44	–	0.53	0.71	0.69	0.85	0.74	en	0.44	0.41	–	0.04	0.62	0.26	0.36	0.31
fi	0.91	0.73	0.82	–	0.78	0.66	0.76	0.75	fi	0.05	0.06	0.04	–	0.00	0.03	0.06	0.02
fr	0.48	0.22	0.64	0.00	–	0.79	0.55	0.81	fr	0.26	0.27	0.62	0.00	–	0.28	0.20	0.25
it	0.53	0.31	0.68	0.32	0.8	–	0.62	0.92	it	0.17	0.22	0.26	0.03	0.28	–	0.13	0.50
sv	1.00	0.54	0.93	0.61	0.71	0.66	–	0.71	sv	1.00	0.32	0.36	0.06	0.20	0.13	–	0.16
pt	0.63	0.38	0.73	0.14	0.89	0.94	0.67	–	pt	0.16	0.18	0.31	0.02	0.25	0.50	0.16	–

Table 6: Pairwise normalized language similarities for all non-`lang2vec` measures in our study.

C BLEU on Validation Set per Training Epoch

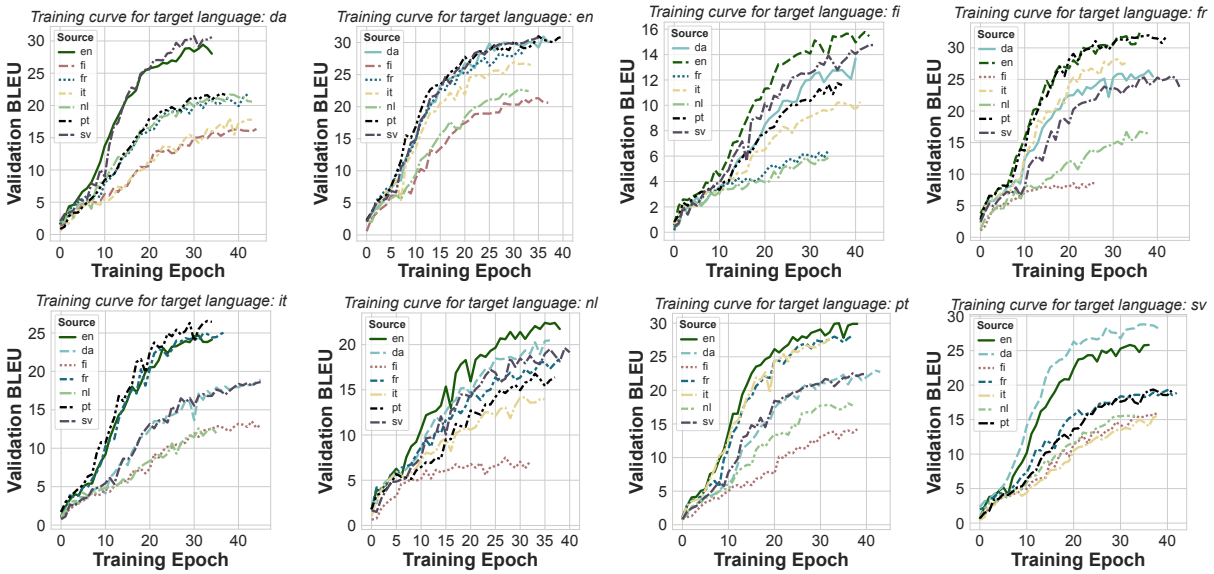


Figure 9: BLEU on validation set per training epoch per target language.

D Loss per Training Epoch per Target Language

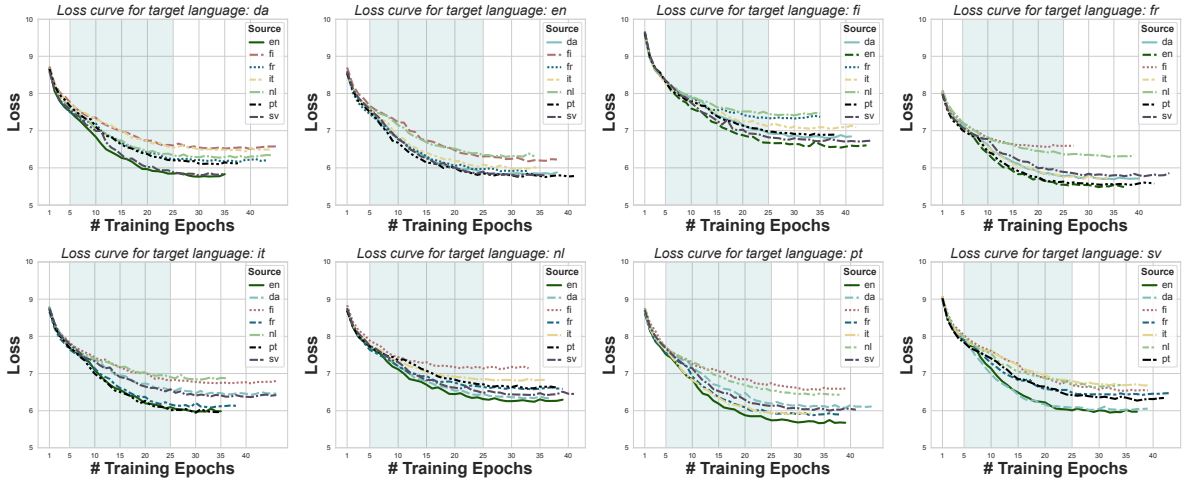


Figure 10: Loss curves per target language in training, with marked the slope that we calculate.

Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets

Tom Kocmi Cohere
Ekaterina Artemova Toloka AI
Eleftherios Avramidis DFKI
Rachel Bawden Inria, Paris, France
Ondřej Bojar Charles University
Konstantin Dranch CustomMT
Anton Dvorkovich Dubformer
Sergey Dukanov Dubformer
Mark Fishel University of Tartu
Markus Freitag Google
Thamme Gowda Microsoft
Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh
Marzena Karpinska Microsoft
Philipp Koehn Johns Hopkins University
Howard Lakougna Gates Foundation
Jessica M. Lundin Gates Foundation
Christof Monz University of Amsterdam
Kenton Murray JHU
Masaaki Nagata NTT
Stefano Perrella Sapienza University of Rome
Lorenzo Proietti Sapienza University of Rome
Martin Popel Charles University
Maja Popović DCU & IU
Parker Riley Google
Mariya Shmatova Toloka AI
Steinþór Steingrímsson The Árni Magnússon Institute
Lisa Yankovskaya University of Tartu
Vilém Zouhar ETH Zurich

Abstract

This paper presents the results of the General Machine Translation Task organized as part of the 2025 Conference on Machine Translation (WMT). Participants were invited to build systems for any of the 30 language pairs. For half of these pairs, we conducted a human evaluation on test sets spanning four to five different domains. We evaluated 60 systems in total: 36 submitted by participants and 24 consisting of translations we collected from large language models (LLMs) and popular online translation providers. This year, we focused on creating challenging test sets by developing a difficulty sampling technique and using more complex source data. We evaluated system outputs with professional annotators using the Error Span Annotation (ESA) protocol, except for two language pairs, for which we used Multidimensional Quality Metrics (MQM) instead. We continued the trend of increasingly shifting towards document-level translation, providing the source texts as whole documents containing multiple paragraphs.

1 Introduction

The Tenth Conference on Machine Translation (WMT25)[1] was held in conjunction with the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025). This 20th iteration of the conference builds on previous editions (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023, 2024a)[2] and hosted ten shared tasks on various aspects of machine translation (MT). This paper describes one of these tasks: the General Machine Translation shared task.

The goal of the General Machine Translation shared task is to explore the translation capabilities of current systems across a broad range of languages. Determining how best to test general MT performance is a research question in itself. Numerous phenomena could be evaluated, the most important of which include:

- **different domains** (news, medicine, IT, patents, legal, social, gaming, etc.),
- **styles of text** (formal or spoken language, fiction, technical reports, etc.),
- **non-standard user-generated content** (errors, code-switching, abbreviations, etc.),
- **source modalities** (text, speech, image).

[1] www2.statmt.org/wmt25/

[2] WMT was organized as a workshop for ten years (2006-2015) before becoming a conference, making this the 20th event overall.

Proceedings of the Tenth Conference on Machine Translation, pages 355–413
November 8-9, 2025

Since individually evaluating all of these phenomena is infeasible, we focus on a selection of domains: news, social/user-generated content, speech, literary, and educational texts. These domains were chosen to cover diverse styles while remaining broadly accessible; thus allowing for evaluation by human annotators without in-domain expertise. However, due to limited access to monolingual data, the set of evaluated domains is not identical across all language pairs.

Our evaluation includes the following 16 language pairs, with those new to this year's task marked by ***(new)***:

Czech→Ukrainian,
Czech→German,
Japanese→Simplified Chinese,
English→Egyptian Arabic ***(new)***,
English→Bhojpuri ***(new)***,
English→Simplified Chinese,
English→Czech,
English→Estonian ***(new)***,
English→Icelandic,
English→Italian ***(new)***,
English→Japanese,
English→Korean ***(new)***,
English→Maasai, Kenya ***(new)***,
English→Russian,
English→Serbian, Cyrillic script ***(new)***,
English→Ukrainian.

A new multilingual subtrack, subject only to automatic evaluation, also introduced 15 additional language pairs:

English→Bengali,
English→German,
English→Greek,
English→Hindi,
English→Indonesian,
English→Kannada,
English→Lithuanian,
English→Marathi,
English→Persian,
English→Romanian,
English→Serbian, Latin script,
English→Swedish,
English→Thai,
English→Turkish,
English→Vietnamese.

Furthermore, this year's task is also distinguished by several key choices:

- **Non-textual modalities:** In addition to textual data, we also incorporated audio and image sources. For the speech domain, participants received audio files with their automatic speech recognition (ASR) transcriptions. For the social domain, screenshots of posts were provided. Participants had the option to use the original audio directly, rather than relying on the provided ASR text.
- **Difficulty sampling:** We used the difficulty sampling method to select more challenging documents, hence increasing the overall difficulty of the test set.
- **Evaluation:** For most languages, we use the Error Span Annotation protocol (ESA; Kocmi et al., 2024b) which combines aspects of DA (Graham et al., 2013) and MQM (Lommel et al., 2014a). For English→Korean and Japanese→Chinese, we use the MQM annotation schema instead.
- **Document-level test set:** Each source text is an entire document (e.g., a news article or social media thread),[3] which is then segmented while preserving paragraph boundaries. This allows us to evaluate translations within their full document context, while giving participants the flexibility to choose their translation strategy: processing the entire document at once, or splitting it by segments or paragraphs.[4]
- **Training corpora:** We prepared a list of recommended training corpora, adding document-level information and CometKiwi22 (Rei et al., 2022) scores for most data sets.

Finally, as in previous years, this year's shared task included several test suites that focused on a range of **translation challenges**, described in Section 8.

All submissions to the General MT Task, along with sources, references, and human judgments, are available in the dedicated GitHub repository.[5]

This paper is organized as follows. We first describe the data used in the shared task, detailing the collection and preparation of our test sets (**Section 2**) and outlining the permitted training data for the constrained track (**Section 3**). Next, we introduce the participating systems, including the large language model baselines added by the organizers (**Section 4**). We then explain our evaluation methodology, covering both automatic metrics (**Section 5**) and human evaluation protocols (**Section 6**). Finally, we present the official results

[3] In some cases, an initial section of a document was used rather than the full text.

[4] No sentence-level segmentation is provided.

[5] github.com/wmt-conference/wmt25-general-mt

(**Section 7**), describe the test suites (**Section 8**), and offer concluding remarks (**Section 9**).

Findings of the General MT Task. We make the following observations:

★ **The number of participants continues to grow:** Participation increased again this year, with a total of **36 submissions**. The majority of these participants used LLMs in their systems, most commonly by fine-tuning (Section 4).

★ **Automatic scores are biased:** Although Shy-hunyuan-MT placed first for all but one language pair in the automatic rankings, human evaluation revealed its performance was considerably lower than that of the top-rated systems (Section 7.2).

★ **Human translations are not always in the winning cluster:** Human references are in the winning cluster for only six out of 15 language pairs (Section 7.2).

★ **Constrained models challenge the performance of LLMs:** The top-performing constrained system was Shy-hunyuan-MT, which placed in the winning cluster for 11 language pairs within its category followed by Algharb placed in winning cluster for 6 language pairs. The best system overall, was Gemini 2.5 Pro, which was in top cluster in 14 language pairs (Section 7.2).

★ **Speech domain was the most challenging:** The speech domain texts were most challenging to translate (likely due to ASR errors) while literary texts were the easiest (Section 7.2).

★ **SOTA systems still struggle with robustness:** Analysis from six specialized test suites reveals that state-of-the-art (SOTA) systems still struggle with robustness to non-standard input, linguistic complexity, domain-specific terminology and gender choice/agreement in particular language pairs. This is despite notable improvements from advanced LLMs in areas like inclusivity and performance in certain specialized domains (Section 8).

2 Test Data

In this section, we describe the test data collection process (Section 2.1) and the creation of human reference translations (Section 2.2). This year, we introduced a new difficulty-based sub-sampling of source texts procedure (Sections 2.1.1 and 6). Motivated by the ever-increasing capabilities of modern MT systems, this step is designed to make our test sets more challenging by selecting source documents that are estimated to be more difficult to translate.

2.1 Collecting test data

Collecting source data. As in previous years, the test sets consist of unseen translations created specifically for the shared task and released publicly as translation benchmarks. We collected public domain or open-licensed source data from a range of domains, focusing on the most recent data available to minimize potential overlap with the pretraining and fine-tuning data of the systems under evaluation. Importantly, for all language pairs, the source texts were originally written in the source language and subsequently translated into the target languages by human translators. This approach is crucial to avoid "translationese" in the source texts, which can negatively affect evaluation accuracy (Toral et al., 2018; Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020).

Domain and language coverage. We collected data from six domains (news, literary, speech, social, educational, and dialogue) and three source languages (Czech, English, and Japanese). However, not all domains cover all three source languages. Detailed statistics on our test sets, including the language coverage for each domain, are provided in Table 1.

2.1.1 Difficulty sampling

To identify documents that are particularly challenging for modern MT systems, we adopted the translation difficulty estimation introduced in Proietti et al. (2025). Specifically, we estimated the difficulty of the collected source documents using their best-performing estimator, `sentinel-src-25`.[6] This model is an improved version of `sentinel-src`, a regression model, based on XLM-Roberta Large (Conneau et al., 2020), which was trained to predict translation quality using only the source text (Perrella et al., 2024).

For each document, we estimated the difficulty of each paragraph and then averaged the results to derive a document-level difficulty score. We then retained the most difficult documents for inclusion in our test sets. To ensure this process did not introduce ill-formed or garbled sources, we also

[6] huggingface.co/collections/Prosho/translation-difficulty-estimators-6816665c008e1d22426eb6c4

Language pair	News	Literary	Speech	Social	Education	Dialogue
			#words			
English→*	8,917	9,921	9,007	8,925	-	9,297
Japanese→Chinese	10,427	12,121	10,655	10,589	-	-
Czech→*	8,643	-	8,898	9,311	9,022	7,660
			#segments (#words/segment)			
English→*	94 (94.86)	85 (116.72)	62 (145.27)	91 (98.08)	-	52 (178.79)
Japanese→Chinese	93 (112.12)	74 (163.80)	59 (180.59)	108 (98.05)	-	-
Czech→*	132 (65.48)	-	86 (103.47)	99 (94.05)	83 (108.70)	52 (147.31)
			#documents (#segments/document)			
English→*	14 (6.71)	2 (42.50)	62 (1.00)	9 (10.11)	-	26 (2.00)
Japanese→Chinese	32 (2.91)	2 (37.00)	59 (1.00)	13 (8.31)	-	-
Czech→*	38 (3.47)	-	86 (1.00)	48 (2.06)	56 (1.48)	26 (2.00)
			#sentences (#sentences/segment)			
English→*	364 (3.87)	873 (10.27)	605 (9.76)	572 (6.29)	-	837 (16.10)
Japanese→Chinese	363 (3.90)	305 (4.12)	712 (12.07)	343 (3.18)	-	-
Czech→*	497 (3.77)	-	556 (6.47)	805 (8.13)	838 (10.10)	1017 (19.56)
			Type-token ratio of source texts			
English→*	0.41	0.30	0.24	0.36	-	0.14
Japanese→Chinese	0.27	0.17	0.17	0.24	-	-
Czech→*	0.52	-	0.42	0.44	0.51	0.23

Table 1: Per-domain statistics of our test sets, calculated on the source side. To compute word counts, we used simple whitespace tokenization for Czech and English and the MeCab morphological analyzer for Japanese. Sentence segmentation was carried out using the language-specific Spacy models for English and Japanese, and the multilingual Spacy model for Czech (Honnibal and Montani, 2017), given that a language-specific one is not available.

manually validated the selected texts and discarded any problematic ones.

We applied this difficulty-based sub-sampling only to the news, speech, and social domains, as the amount of source data available for the other domains was insufficient for the procedure.

2.1.2 Test sets statistics

Source text segmentation. To balance the evaluation across domains and source languages, we aimed for a consistent size for each test set, that is, approximately 9,000 words distributed across 60 to 100 segments (see Table 1).[7] We also aimed to keep the average segment length around 100 words, whenever possible. This design enables micro-averaging of results without any single category disproportionately influencing the final scores.

These choices were motivated by the aim to keep approximately the same number of segments per domain, which is important for balancing the domains for manual evaluation. The number of segments per domain is therefore more balanced than in last year's test sets, as can be shown in Table 1. The composition of the texts included remains domain-specific due to the differing natures of the texts. For example, in the literary domain texts are longer and therefore only 2 documents were used for each source language. These were split into a large number of segments (42.5 per document on average for English→*). In contrast, the speech domain, where 59-86 documents were used depending on the source language, with each document forming its own segment. The longest segments are in the speech domain, while the shortest are in the news domain (respectively 145.27 and 94.86 words per segment on average for English→*).

Language-specific adjustments. For practical reasons, we were not always able to meet the objective of having a minimum of 100 words per segment. Most notably, the average segment length was longer in the speech domain for English (145.27 words) and Japanese (180.59 tokens), and in the dialogue domain for English (178.8 words)

[7] A segment contains one or more paragraphs, where each paragraph is defined by a line break in the source document. We then separate the segments from each other using double line breaks.

and Czech (147.3 words).

Finally, for Japanese, based on the 1-to-2 ratio derived from our observations of previous WMT Japanese-English test sets, we targeted approximately 18,000 characters per domain. We adopted a character-based metric, because the lack of spaces in Japanese complicates word segmentation. This approach also aligns with industry standards for translation pricing.

In the following paragraphs, we provide other information regarding the data collection process, but specific to each domain.

News domain. For this domain, the data was prepared similarly to previous years (Kocmi et al., 2024a). We collected news articles published in February 2025 on online news sites and extracted the text while preserving the paragraph boundaries. This year we specifically aimed to extract articles that were more challenging to translate. For English, we limited the news crawl to opinion pieces on the basis that they tend to use a more complex, literary style than the straightforward event reporting. For Czech and Japanese, we extracted a larger pool of news covering the entire month to follow up with down-sampling.

Social domain. The social domain data was sourced using the Mastodon Social API.[8] Mastodon is a federated social network that is compatible with the W3C standard ActivityPub (Webber et al., 2018). Users publish short-form content known as "toots", with the possibility of replying to other toots to form threads. We decided to use the original server, `mastodon.social` because of its large community and publicly available toots.

We collected data in March and April of 2025, using the reported language ID label to target the source languages of interest.

Given the document-level nature of the task, our aim was to collect threads comprising multiple toots. Our sourcing therefore involved regularly scraping random toots from the previous hour but also identifying and scraping any missing toots that made up threads only partially sourced (identified using the 'in_reply_to_id' attribute of already sourced toots). To avoid spam and uninformative toots, we removed empty toots, texts that appeared several times (probable spam), texts from accounts that produced a large number of toots overall (we set this to 100) and from accounts we heuristically identified as being news outlets or bots (containing the keywords 'bot', 'news', 'weather', 'sports', 'feeds' or 'press' in their handle, as well as a few known media accounts). We created threads from the individual toots and manually selected threads of interest from threads of minimum 2 and maximum 100 toots. Our selection was based on having a diverse range of topics and targeting those characteristic of non-standard user-generated content.

The selected documents contain either a whole toot or a line of text within a toot (depending on whether the toot contained newlines, i.e. there is no distinction between newlines indicating a boundary between two toots and a newline within a toot). Each segment can therefore contain one or several sentences, depending on the original composition of the toots.

A new aspect of the task this year was the inclusion of screenshots capturing entire conversations. This decision was influenced by requests from human translators, researcher efforts to expand last year's test set (Deutsch et al., 2025), the evolving nature of social media, and an overall interest in exploring the impact of multimodal translation. To accomplish this, we focused on Mastodon conversations that included an image. While the image is not needed for translation, our objective was to provide visual context for translators if needed. Furthermore, even without an image, non-standard domains like social media often convey meaning through layout features such as whitespace, text positioning, and other non-textual elements that may be lost without visual reference. After filtering, we had 481 conversations containing an image for final selection. See an example of the screenshots in Figure 1.

Due to insufficient data for Japanese in the social domain, we adjusted the size of the news test set to compensate, targeting approximately 24,000 characters.

Similarly, we were unable to obtain a sufficient amount of Czech Mastodon data. We therefore decided to use personal communication (usually between a Czech and Ukrainian speaker) from Charles Translator[9] for the Czech Social domain. This data is similar to the domain called "Personal" last year. The texts were collected with users' opt-in consent, filtered and pseudonymized in the same way as in the last three years (Kocmi et al., 2022). Each document is one conversation with one user.

[8] mastodon.social/api/v1/timelines/public

[9] translator.cuni.cz (Popel et al., 2024)

The lines reflect the formatting provided by the users. Segment boundaries (empty lines) were added based on the content (trying not to split a paragraph or sentences that are closely related), so as a result the average number of words per segments is close to 100.

Literary Domain. For the English source texts, we selected an amateur-written story from the Archive of Our Own[10] based on several criteria: its recency (published April 2025), its narrative quality, and the absence of explicit sexual or harmful content. The first two chapters, comprising approximately 9.9k words, were treated as two documents. The documents were then segmented so that each segment contained at least 100 words and all paragraph boundaries were maintained.

For the Japanese source texts, we selected two short stories[11] from Aozora Bunko, a public-domain digital archive of Japanese literature.[12] Selection was guided by three criteria: recent publication on the platform, the use of modern orthography (*shinjitai*), and a prose style accessible to contemporary readers. In line with the methodology for the English texts, both stories were segmented into passages of similar length ensuring that all paragraph-level boundaries were preserved. For the test set, we used all six chapters of the first story (*Bishōjo Ichiban-nori*) and the first four chapters of the second story (*Omokage*).

Speech domain. The speech corpus was compiled from Creative Commons–licensed YouTube videos. For each language, we collected approximately 2,700 videos retrieved with 200 distinct queries spanning documentaries, instructional content, lectures, interviews, news, lifestyle vlogs, sports, and performing arts.

From each video, a segment was randomly sampled, with a minimum duration of 30 seconds and a maximum of 50 seconds, containing at least 30% speech. This constraint was introduced to balance the amount of context required for ASR and translation with the number of available documents.

Transcription of the segments was carried out using the *Whisper-large-v3* model (Radford et al., 2023). For English and Czech, punctuation was expected to be provided directly by Whisper. In cases where the model hallucinated or returned text without punctuation, *Whisper-large-v2* was used instead. For Japanese, Whisper almost never returned punctuation. Therefore, an additional punctuation restoration model was applied.[13]

After the sampling procedure, 90% of the documents were discarded, and subsequent manual filtering retained approximately one third of the remaining examples.

While the shared task participants had access only to the original videos and automatic transcripts, the reference translations from Czech to Ukrainian and German where prepared with an initial manual correction of the ASR errors using the videos as a first step. As a result, the Ukrainian and German translations are expected to be more accurate in cases where the original transcript was ambiguous.

Education domain. The Educational domain includes selected exercises from an online portal *Škola s nadhledem*[14] for elementary-school students from various subjects (chemistry, geography, Czech language, etc.). Some segments are not full sentences but short phrases. The reference translations into Ukrainian and German for this domain were created by professional translators within the EdUKate project. Last year, each page of an exercise was treated as a separate document, while this year, each exercise (with all its pages) was compiled into a single document. To meet the target of 9k words per document and 100 words per segment, we excluded documents with less than 90 words. Longer documents were split into multiple segments along page boundaries to ensure that no segment is longer than 200 words.

Dialogue domain. The Dialogue domain texts originates in the MultiWOZ2.4 dataset (Ye et al., 2022)[15] simulated dialogues between a user and a mock dialogue system (Wizard-of-Oz setup) responding to the user's requests in multiple domains. Each document contains two parts: a description of what should be achieved in a dialogue with the agent (e.g. find a restaurant in Cambridge or finding and booking accommodation), and the dialogue of the user and the agent itself. As such, the sentences in this dataset are rather simple for translation, but the point lies in cross-sentence co-

[10] archiveofourown.org

[11] *Bishōjo Ichiban-nori* ("Pretty Girl, First to Arrive") from 1938 and *Omokage* ("Reminiscence") from 1942 by Yamamoto Shūgorō.

[12] aozora.gr.jp

[13] huggingface.co/1-800-BAD-CODE/punct_cap_seg_47_language

[14] skolasnadhledem.cz

[15] github.com/smartyfh/MultiWOZ2.4

herence and primarily in gender and politeness preservation (or biases), which was promoted in the dataset through our translation process: We selected a subset of the MultiWOZ2.4 test set and professionally translated it from the original English to Czech. English typically keeps the gender of the parties unexpressed and conveys little or no markers of politeness. The first official translation to Czech was thus heavily biased towards the masculine gender and formal politeness level, both of which are explicit in Czech. We requested the translators to add more variance in this regard, i.e. pretend that one or both of the parties are female, and vary the politeness level (formal vs. informal). This varied Czech should not be treated as a canonical reference, but we used it as an interesting *source* for Czech→German translation because German is similarly explicit in gender and politeness as Czech. The participating systems in Czech→German translation thus have to demonstrate their ability to preserve these features (rather than losing them, for example, in implicit or explicit pivoting through English).

2.2 Human References

The test sets were translated by professional translation agencies according to the brief in Appendix B. Since each language pair was sponsored by a different partner, multiple translation agencies contributed, which may account for some variability of the final translations across languages.

Automatic quality assessment of human translations. The quality of human references is critical for reference-based metrics (Freitag et al., 2023). However, obtaining high-quality translations is challenging even with professional translators. This challenge was particularly salient this year, as our difficulty sampling approach (Section 2.1.1) intentionally selected hard-to-translate source texts. Therefore, following WMT24, we report scores from automatic evaluation methods to assess the quality of the collected human references. For this evaluation, we employ the GEMBA-ESA (Kocmi and Federmann, 2023a) as an LLM-as-a-Judge method, using two independent judges: GPT-4.1[16] and Command A (Team, 2025). Our full automatic evaluation approach is detailed in Section 5.

[16] openai.com/index/gpt-4-1/

Discussion on the quality of human references. Table 2 shows the average GEMBA-ESA scores for the human reference translations, broken down by language pair and domain. The two language pairs with the lowest average GEMBA-ESA scores are English→Russian and English→Icelandic. For Russian, this aligns with its human evaluation results, i.e., human translations end up in the third cluster. For Icelandic, however, the pattern diverges: its human reference is the only item in the first cluster, outperforming the best MT system (Gemini 2.5 Pro) by a margin of 20 points (ESA). Given this discrepancy, and the fact that GPT-4.1 largely disagrees with Command A's lower score for the Icelandic reference, it is possible that Command A is systematically underrating this particular translation.[17] This hypothesis is further supported by Command A's training data as out of all target languages included in the evaluation, Command A was not optimized to support Icelandic, Estonian, and Serbian (Team, 2025). Consistent with this, both Estonian and Serbian show notable negative difference values (Command A $<$ GPT-4.1): -12.36 and -7.83, respectively. The main counterexample to this trend is English→Japanese; although Japanese is supported by Command A, it also shows a notable negative difference (-8.91) and, like Icelandic, its human reference also ranks alone in the first cluster. Across the remaining language pairs, Command A and GPT-4.1 yield similar absolute scores, with a mean difference of -4.90 points.

Finally, the English→Bhojpuri and English→Maasai pairs were excluded from this QE evaluation, as metric reliability has not been established for these low-resource languages (Section 5; Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhujan et al., 2025).

2.3 Test Suites

In addition to the test sets of the regular domains, the test sets given to the system participants were augmented with several *test suites*, which are custom-made test sets focusing on particular aspects of MT translation. The test suites were contributed and evaluated by test suite providers as part of a decentralized sub-task, detailed in Section 8. Across all language pairs of the shared task, test suites contributed 72,449 source test segments

[17] Command A scored the Icelandic human reference translation 15.04 points lower than GPT-4.1.

	Lit.	News	Social	Speech	Dial.	Edu	Avg.	Hum.
CS-DE	—	75.9	74.4	72.0	91.3	74.5	77.6	2
CS-UK	—	78.4	79.9	72.3	—	77.2	77.0	2
EN-AR	69.2	66.0	79.0	77.7	—	—	73.0	1
EN-CS	79.0	74.9	78.1	77.1	89.2	—	79.7	3
EN-ET	82.9	75.1	79.2	69.6	—	—	76.7	1
EN-IS	78.1	70.2	75.5	67.9	—	—	72.9	1
EN-JA	83.1	76.7	84.1	80.0	—	—	81.0	1
EN-KO	85.2	82.4	83.9	83.0	—	—	83.6	1
EN-RU	74.3	66.3	75.1	62.9	—	—	69.7	3
EN-SR	84.0	73.3	81.9	76.4	—	—	78.9	4
EN-UK	85.0	82.4	85.4	83.3	—	—	84.0	2
EN-ZH	79.2	68.8	78.0	72.4	—	—	74.6	2
JA-ZH	79.9	82.7	86.6	72.6	—	—	80.5	1

Table 2: GEMBA-ESA scores for human references. Each domain cell is the arithmetic mean of Command A and GPT-4.1; the Avg. column reports the macro-average across available domains. The last column is the human cluster assigned using the ESA protocol.

(detailed numbers can be found in Table 14).

3 Training Data

Similar to the previous years, we provide a selection of parallel and monolingual corpora for model training. The provenance and statistics of the selected parallel datasets are provided in the appendix in Table 18 and Table 19. Specifically, our parallel data selection include large multilingual corpora such as Europarl-v10 (Koehn, 2005), Paracrawl-v9 (Bañón et al., 2020), CommonCrawl, NewsCommentary-v18.1, WikiTitles-v3, WikiMatrix (Schwenk et al., 2021), TildeCorpus (Rozis and Skadiņš, 2017), OPUS (Tiedemann, 2012), CCAligned (El-Kishky et al., 2020), UN Parallel Corpus (Ziemski et al., 2016), and language-specific corpora such as YandexCorpus,[18] ELRC EU Acts, JParaCrawl (Morishita et al., 2020), Japanese-English Subtitle Corpus (Pryzant et al., 2018), KFTT(Neubig, 2011), TED (Cettolo et al., 2012), and back-translated news.

Links for downloading these datasets were provided on the task web page. We have automated the data preparation pipeline using MTDATA (Gowda et al., 2021).[19] This year's monolingual data include the following: News Crawl, News Discussions, News Commentary, CommonCrawl, Europarl-v10 (Koehn, 2005), Extended CommonCrawl (Conneau et al., 2020), Leipzig Corpora (Goldhahn et al., 2012), UberText and Legal Ukrainian.

[18] github.com/mashashma/WMT2022-data
[19] statmt.org/wmt25/mtdata

Our automated dataset preparation pipeline made a best-effort attempt to preserve document identifiers whenever available in the source. In addition, we have precomputed and shared a quality estimation (QE) metric scores on training data to facilitate data filtering. Specifically, we shared COMETKIWI22 (Rei et al., 2022) annotations for all parallel segments in the training corpora. These were computed using the fast and efficient PYMARIAN framework (Gowda et al., 2024), which enabled QE scoring at scale.

4 System Submissions

This year, we received 96 submissions from 36 participating teams. While the number of submissions is slightly lower than last year's 105, the number of participating teams increased by roughly a third.

In line with previous years, we included translations from three public MT services, anonymized as ONLINE-{B,G,W}. We also added contrastive translations from 20 LLMs—including commercial products like GPT-4.1 and open weights models like Llama3.1—and one encoder-decoder system (NLLB). This brought the total to 60 participants.

All participating systems are listed in Table 3. A more detailed description of each submitted system is included in Appendix C, as provided by the authors at the submission time. Section 4.1 discusses the general trends in chosen architectures and training strategies. Section 4.2 presents details on LLM benchmark usage in the task. Section 4.3 outlines two tracks, constrained or unconstrained, to which participants could submit outputs. Section 4.4 describes the submission system platform.

4.1 Architectures and Strategies

Each participating team was asked to submit a form detailing their approach along with an optional description paper. This section discusses the submissions with a summary of strategic details, such as the approach or base model, is provided in Table 3.

Architectures. This year generative language models (denoted as LLM in the table) were the predominant approach, used by all but one external submission. Still, six external team and one organizer submission mentioned using the encoder-decoder architecture, and several others used hybrid approaches.

Base models. The most popular pretrained language models were Qwen (7 submissions), Eu-

Team	Approach	Model	Size	Data	Training	Post-proc	Best rank
Algharb	LLM	Qwen 3	14B	—	SFT, CO	MBR, QE	1-1 (en-zh)
AMI	LLM	Llama 3.2	3B	B, S	CPT, SFT, CO	MBR, QE, rules	11-11 (en-is)
COILD-BHO	LLM	Llama 2	7B	B	CPT, SFT, CO	APE	13-15 (en-bho)
CommandA-WMT	LLM	CommandA	111B	E, S	CO	MBR, reason	5-6 (en-cs)
CUNI-Doc Transformer	enc-dec	from scratch	<1B	B	CPT	checkpoint avg	—
CUNI-EdUKate-v1	LLM	EuroLLM it	9B	B, E, S	SFT, CO, Ada	—	—
CUNI-SFT	LLM	EuroLLM	9B	B, E	SFT	—	16-17 (en-sr)
CUNI-Transformer	enc-dec	from scratch	<1B	B, S	CPT	checkpoint avg	—
DLUT_GTCOM	LLM, enc-dec	Gemma 3	27B	B, E, S	CPT, SFT	prompt	10-15 (en-sr)
HYT	LLM	Hunyuan-TurboS	—	B, E, S	CPT, RL	prompt	—
GemTrans	LLM	Gemma 3	27B	S	SFT, RL	APE	1-4 (en-it)
In2x	LLM	Qwen 2.5	72B	—	CPT	MBR, ensemble	9-16 (jp-zh)
NTTSU	LLM	Qwen 3	14B	B, S	CPT, SFT, CO	MBR	14-15 (jp-zh)
SalamandraTA	LLM	Salamandra	2B, 7B	B, E	CPT, SFT	MBR, TRR	11-15 (en-sr)
SH	LLM	DeepSeek-R1-Distill-Qwen-Japanese	14B	—	SFT, CO	MBR, APE CO	—
Shy-hunyuan-MT	LLM	Hunyuan	7B	B, E, S	SFT, CO	prompt	1-3 (en-ko)
Systran	LLM	EuroLLM	9B	B, E	CPT, SFT	MBR, QE, ensemble, prompt	13-16 (en-jp)
TranssionMT	enc-dec, LLM	NLLB, EuroLLM	1B, 9B	B, E, S	CPT, SFT	ensemble	9-13 (en-mas)
UvA-MT	LLM	Gemma 3	12B	S	SFT	—	5-10 (en-zh)
Wenyiil	LLM	Qwen 3	14B	B, E, S	SFT, CO	MBR, QE, ensemble, APE	1-3 (en-uk)
Yandex	LLM	YandexGPT	—	B, E, S	CPT, SFT, CO	—	8-10 (en-ru)
Yolu	LLM	Qwen 3	14B	B, E, S	CPT, SFT, CO	MBR, APE, prompt	7-8 (en-et)
CUNI-MH-v2	LLM	EuroLLM it	9B	E, S	Ada, CO	—	16-16 (en-cs)
KYUoM	enc-dec	NLLB	600M	—	Ada	QE	—
Laniqo	LLM	EuroLLM it	9B	S	Ada/CO	MBR, QE, rules	12-17 (en-et)
CGFOKUS	LLM	Qwen 3	235B	—	—	prompt	—
Erlendur	LLM, hybrid	Claude 3.5 Sonnet	—	E	—	APE, prompt	3-4 (en-is)
IR-MultiagentMT	LLM	GPT-4o-mini	—	E	—	prompt	—
IRB-MT	LLM MAS	Gemma 3 it	12B	—	—	reason, prompt	7-7 (en-arz)
Kaze-MT	LLM	Qwen 2.5	72B	—	—	QE, ensemble	—
KIKIS	LLM	Plamo-2-translate	18B	B	—	reason, ensemble, prompt	13-16 (en-jp)
RuZH-Eole	LLM + Estimator	TowerPlus	9B	—	—	QE	17-18 (en-zh)
SRPOL	LLM, hybrid	EuroLLM, NLLB	9B, 3B	—	—	QE, ensemble	12-15 (en-et)

bb88, ctpc_nlp, TranssionTranslate: no information provided, no paper submitted

SYSTEMS ADDED BY THE ORGANIZERS: all LLMs, except NLLB (enc-dec):

Model	Size	Best rank	Model	Size	Best rank	Model	Size	Best rank
AyaExpanse	8B	4-6 (en-mas)	Gemini 2.5 Pro	—	1-1 (many)	NLLB	3.3B	8-10 (en-bho)
AyaExpanse	32B	7-13 (en-mas)	Gemma 3	12B	9-13 (en-mas)	Qwen 2.5	7B	9-13 (en-mas)
Claude4	—	2-4 (cs-de)	Gemma 3	27B	6-10 (cs-uk)	Qwen 3	235B	6-11 (en-zh)
CommandA	111B	3-3 (en-arz)	GPT-4.1	—	1-3 (cs-uk)	TowerPlus	9B	6-6 (en-is)
CommandR	7B	11-14 (en-arz)	Llama 3.1	8B	9-13 (en-mas)	TowerPlus	72B	8-10 (en-is)
DeepSeek V3	671B	2-6 (cs-de)	Llama-4-Maverick	400B	4-5 (en-mas)	ONLINE-B	—	3-4 (en-sr)
EuroLLM	9B	14-16 (en-it)	Mistral	7B	—	ONLINE-G	—	—
EuroLLM	22B	13-17 (en-et)	Mistral-Medium	—	2-5 (en-jp)	ONLINE-W	—	—

Table 3: Submissions to the General MT shared task, including the externally contributed submissions as well as the systems added by the organizers. Row coloring shows unconstrained-track (dark gray) and constrained-track (white) submissions. Entries are ordered lexicographically, with first the submissions that modified the foundation models somehow (training, tuning, etc), then submissions that created adapters without modifying the models and finally the submissions that used models as is. The last column shows the best rank achieved by the submission, as defined in the official results (see Section 7.4) and the translation direction where the rank was achieved.
Abbreviations: **LLM** (decoder-only language model), **enc-dec** (encoder-decoder), **B** (basic data preproc), **E** (elaborate data preproc), **S** (synthetic data), **CPT** (continued pre-training), **SFT** (supervised fine-tuning), **CO** (contrastive optimization/preference tuning), **Ada** (adapters), **MBR** (Minimum Bayes Risk decoding), **QE** (quality estimation), **rules** (rule-based post-processing/regular expressions), **reason** (reasoning in LRMs), **prompt** (prompting), **APE** (automatic post-editing), and **TRR** (translation re-ranking).

roLLM (6 submissions), and Gemma (4 submissions). Twenty-three submissions modified their base model via continued pre-training, supervised fine-tuning, preference optimizations (CPO/DPO), or reinforcement learning (RL). Three submissions trained adapters without changing the model, and eight used prompting without any model training.

Data preparation. For data preparation, 17 submissions reported using basic filtering (e.g., OPUS cleaner or empirical steps), while 15 reported more elaborate techniques (e.g., filtering with quality estimation by utilizing LLM-as-a-judge or CometKiwi). Synthetic data generation, such as back-translation, was reported by 16 teams.

Post-editing. Finally, for post-editing, 16 submissions reported using LLMs (prompt engineering) for automatic post-editing and/or reasoning. Eleven teams reported using Minimum Bayes Risk (MBR) decoding, and nine mentioned using quality estimation separately.

4.2 LLM Benchmark

LLMs have become popular tools for machine translation (Ataman et al., 2025; Chatterji et al., 2025). Following last year, we provide a systematic and unbiased evaluation of the most popular language models on our blind test sets.

Evaluated models. When deciding which LLMs to evaluate, we selected the best performing constrained and best performing unconstrained model from each popular LLM family. In addition, we collected three popular translation provider services as in previous years. The final list of systems is presented in Table 5.

Prompting LLMs for translation. We designed a unified script to collect translations from all LLMs in an identical setup. We used a zero-shot, instruction-following approach, translating full documents at once. To ensure deterministic outputs, we set the temperature to zero. If a model failed to translate a full document while preserving paragraph breaks, we segmented it into paragraphs and translated each one separately.[20] This generic setup may disadvantage systems tuned for specific MT instructions, such as TowerLLM or EuroLLM; these are marked with [M].

[20]The code for collecting translations is available at: github.com/wmt-conference/wmt-collect-translations

Language model	Doc-lvl	Tokens (in/out)	Cost ($)
Gemini-2.5-Pro†	95.1%	2.1 / 16.4 M	250.8
Claude-4	67.5%	2.6 / 3.5 M	60.4
CommandA	70.6%	2.3 / 3.0 M	35.3
GPT-4.1	97.1%	2.0 / 3.5 M	31.7
DeepSeek-V3	57.2%	2.5 / 2.8 M	6.6
AyaExpanse-32B	54.9%	2.5 / 3.0 M	5.7
AyaExpanse-8B	43.2%	2.6 / 2.8 M	5.5
Mistral-Medium‡	54.1%	2.2 / 2.0 M	5.0
Qwen3-235B	65.6%	2.5 / 3.4 M	2.5
Llama-4-Maverick	70.4%	2.3 / 2.2 M	2.5
Qwen2.5-7B	35.7%	3.2 / 3.5 M	2.0
Mistral-7B	35.1%	1.8 / 3.0 M	1.2
Llama-3.1-8B	27.1%	3.3 / 3.1 M	1.2
CommandR7B	49.3%	2.7 / 3.9 M	0.7

Table 4: Ratio of document-level translated data. Token counts are in millions. †Gemini-2.5-Pro used reasoning, increasing cost. ‡Mistral-Medium did not translate four language pairs. Pricing for open-weight models is estimated via together.ai.

Supported languages. We collected translations for all language directions and tried to collect information about which languages are supported and which are not by looking into the original technical reports to see which languages are mentioned.

As shown in Table 4 in column *Doc-lvl*, one of the key limitations of current LLMs is failure to translate a document at once. This is caused by their window size or a failure to keep paragraph breaks.

API inference and cost. We collect all translations via the respective service APIs during the submission period. Table 4 shows the number of input and output tokens as determined by each model's tokenizers. The estimated costs shown are for the main test set and do not include the test suites. While we disabled the "reasoning" mode for Qwen3-235B to prevent collection errors, we did not disable it for Gemini 2.5 Pro, which significantly increased its translation cost.

4.3 Constrained and Unconstrained Tracks

To promote fair comparison and encourage replicability, the WMT25 General MT Task distinguished between two tracks: *Constrained* and *Unconstrained*. These tracks differ in terms of model size, licensing, and reproducibility requirements.

Constrained Track: Systems in this track must adhere to the following criteria:

- Use only models and training data that are publicly available under open-source licenses per-

mitting unrestricted non-commercial use (e.g., Apache, MIT).
- The final model must not exceed 20 billion parameters. Larger models may be used during intermediate stages (e.g., for distillation), but the submitted outputs must be produced by a final model that complies with the size limit.
- Model weights must be released under an open-source license before the camera-ready deadline to ensure replicability.

This track is designed to foster transparency and reproducibility, allowing other research groups to replicate and build upon submitted systems.

Unconstrained Track: This track imposes no restrictions on model size, training data, or licensing. It includes systems built with proprietary tools, closed-source models, or undisclosed training pipelines, such as commercial LLMs (e.g., GPT-based systems). While participation in this track is open, systems are expected to provide as much detail as possible about their setup to support interpretability.

Although this year, we did not restrict training data for either track, we curated a set of corpora that covers the majority of publicly available resources to support participants in building competitive systems.[21]

To assist participants in the constrained track, we provided a non-exhaustive list of suggested models under the 20B parameter limit, including: textual models: Aya Expanse 8B, Aya 101 (13B), Cohere R 7B, LLaMA 7B and 13B, Qwen 2.5 7B, Mistral 7B and 8B, EuroLLM, NLLB; and multimodal models: Whisper, Seamless M4T. The complete list of systems is presented in Table 5.

Systems were evaluated within their respective tracks: constrained systems were compared only against other constrained systems, while unconstrained systems were evaluated in a broader context that includes all submissions.

4.4 System Submission Platform

We used the open-source OCELoT platform[22] to collect system submissions again this year. Given that not all submissions could be included in the human evaluation due to resource constraints, we did not require pre-verification of participating teams. This allowed broader participation and flexibility in the submission process.

[21] www2.statmt.org/wmt25/mtdata

[22] github.com/AppraiseDev/OCELoT

	Model	# Params	Open?
Constrained systems	AyaExpanse-8B	8B	✓
	CommandR7B	7B	✓
	EuroLLM-9B	9B	✓
	Gemma-3-12B	12B	✓
	Llama-3.1-8B	8B	✓
	Mistral-7B	7.3B	✓
	NLLB (NLLB-200)	3.3B	✓
	Qwen2.5-7B	7.6B	✓
	TowerPlus-9B	9B	✓
Unconstrained systems	AyaExpanse-32B	32B	✓
	Claude-4	—	✗
	CommandA	111B	✓
	DeepSeek-V3	671B (37B act.)	✓
	EuroLLM-22B	22B (preview)	✓
	Gemma-3-27B	27B	✓
	Gemini-2.5-Pro	—	✗
	GPT-4.1	—	✗
	Llama-4-Maverick	—	✓
	Mistral-Medium	—	✗
	ONLINE-B	—	✗
	ONLINE-G	—	✗
	ONLINE-W	—	✗
	Qwen3-235B	235B (22B act.)	✓
	TowerPlus-72B	72B	✓

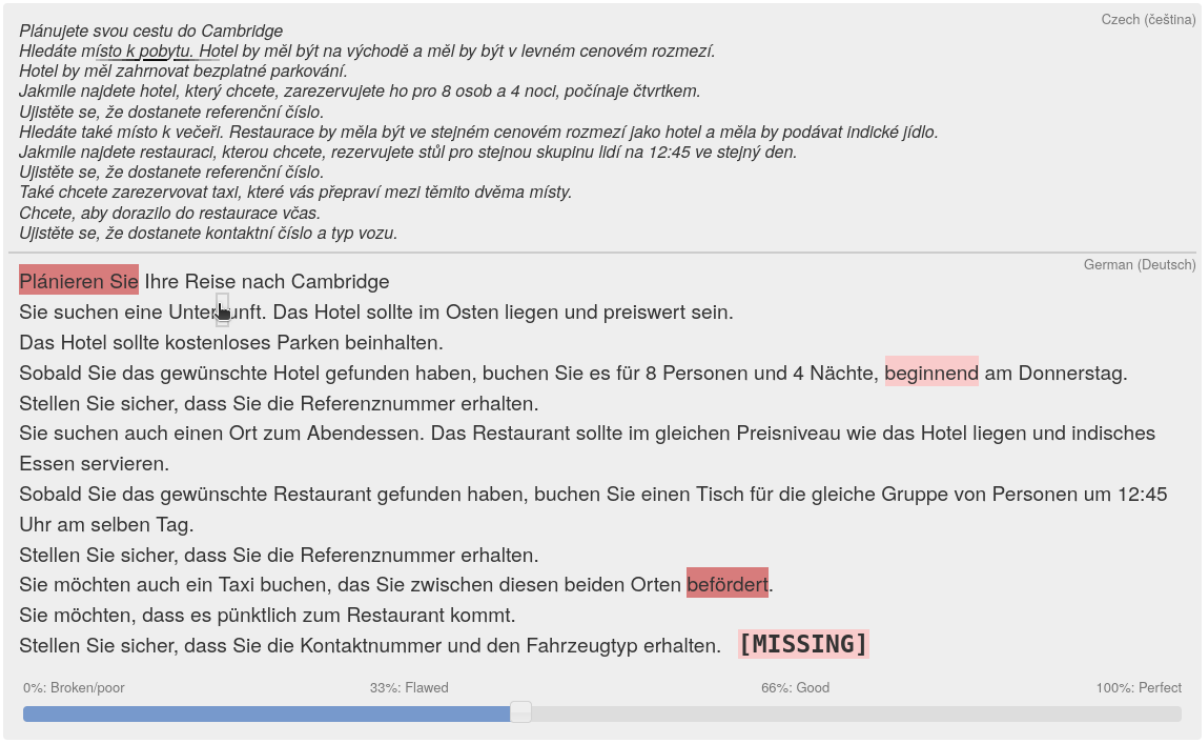
Table 5: List of constrained and unconstrained systems with parameter count. Open-weight models were marked with a cmark (✓).

Participants were asked to submit their systems in a single JSONL file, covering all language pairs, and to use a provided verification script to ensure that the submission adheres to strict formatting and completeness requirements. The verification process included the following checks:

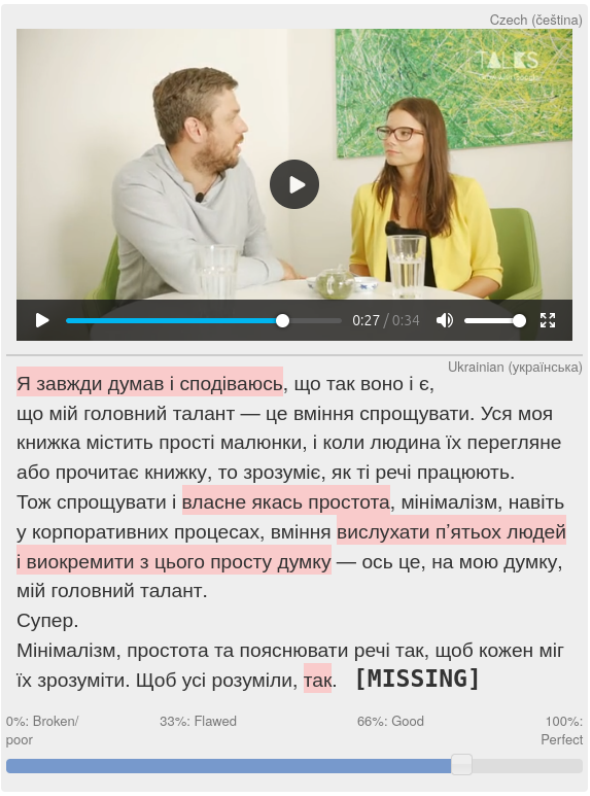
- **Completeness of translations:** while participants could submit outputs for only a subset of languages, each submitted language had to be fully translated.
- **Format verification:** ensured that all documents from the testset were translated with correct paragraph boundaries preserved.
- **Inclusion of testsuite translations:** verified that testsuite segments were not omitted because of their length or other reasons.

To avoid premature publication of rankings based on automatic metrics, all submissions were displayed anonymously on the leaderboards during the submission period.

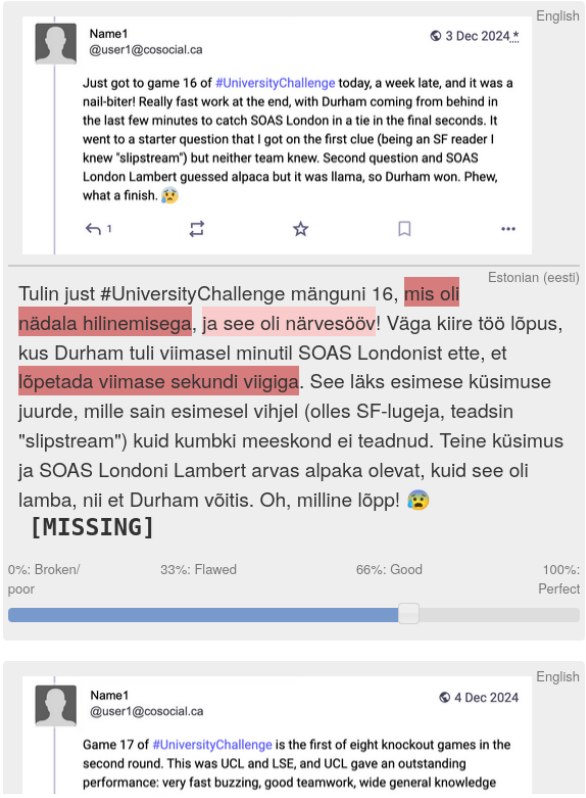
Teams had one week after the submission deadline to select a single primary submission, specify the track for that submission, and submit an abstract for their system description paper. These steps were mandatory for a system to be included in the human evaluation campaign.



(a) Screenshot of Czech→German annotations in the dialogue domain. Mouseover on the target side shows coarse alignment on the source side.



(b) Screenshot of Czech→Ukrainian annotations in the speech domain. The video can be paused and replayed.



(c) Screenshot of English→Estonian annotations in the social domain. The image partially shows the next segment which provides context.

Figure 1: Three screenshots of ESA (Kocmi et al., 2024b) annotations. ESA shows multiple segments within a document at once as well as video or image sources. After marking the individual error spans, the annotator assigns the final segment score from 0 to 100.

Language pairs	Annotators' profile	Tool
English→Chinese,Japanese	Microsoft annotators: bilingual target-language native speakers, professional translators or linguists, experienced in machine translation evaluation.	Appraise ESA
Czech→Ukrainian,German English→Czech	ÚFAL Charles University annotators: linguists, annotators, researchers, and students who were native speakers of one of the languages and high proficiency in the other.	Appraise ESA
English→Ukrainian,Russian Arabic,Serbian Maasai,Bhojpuri	Toloka AI paid expert crowd: Bilingual native target-language speakers who were high-performing on the platform.	Appraise ESA
English→Estonian	Professional translators from Luisa Language Solutions.	Appraise ESA
English→Icelandic	The Árni Magnússon Institute for Icelandic Studies annotators: bilingual target-language native speakers, paid translators with 3–25 years of experience in Icelandic↔English translation.	Appraise ESA
English→Italian	Cohere annotators: professional in-house employees experienced with annotations of general LLM outputs, bilingual speakers native in Italian.	Appraise ESA
English→Korean Japanese→Chinese	Professional translators from Venga Global.	Anthea MQM

Table 6: Annotators' profiles and annotation tools for each language pair in the human evaluation. All annotators were paid a fair wage in their respective countries.

5 Automatic Evaluation

As in the last year, a high number of submissions[23] made full comprehensive manual evaluation infeasible. We therefore employed automatic metrics to select systems for human evaluation via a procedure we call AUTORANK. This year's procedure improves upon WMT24 by incorporating a broader set of metrics and a revised aggregation method.

Metrics. For most language pairs (see "low-resource exception" below) the AUTORANK is a combination of three distinct families of evaluation methods:

- **LLM-as-a-Judge (reference-less).** We use GEMBA-ESA (Kocmi and Federmann, 2023b) with two independent judges: GPT-4.1[24] and Command A (Team, 2025), both in a reference-less setting.
- **Trained Reference-based Metrics.** Two supervised metrics trained to approximate human quality judgments with references: MetricX-24-Hybrid-XL[25] (Juraska et al., 2024) and XCOMET-XL[26] (Guerreiro et al., 2024).
- **Trained Quality Estimation (QE).** The reference-less QE metric CometKiwi-XL[27] (Rei et al., 2023), which is also trained to mimic human judgments.

This combination of reference-based and reference-less (or QE) methods is designed to balance their complementary failure modes. Reference-based metrics typically achieve a higher correlation with human judgments when high-quality references are available, while reference-less methods reduce susceptibility to reference bias when references are suboptimal (Freitag et al., 2023). We also account for known issues with specific metrics. To mitigate a common QE pitfall, i.e., being fooled by fluent output in the wrong language, the GEMBA-ESA prompt explicitly specifies the target language. However, while GEMBA-ESA is intended to reduce bias toward systems that use re-ranking, we note that some participants incorporated it directly as a reward model.

System-level scores. The system-level score for each language pair is the average of its paragraph-level (segment-level) scores from each metric across the testset. In particular, paragraphs constitute the input units for all the metrics. We make one exception for language pairs without human

[23] We received submissions from 36 unique teams. A total of 43 teams initially registered, but 7 later withdrew or were disqualified.

[24] openai.com/index/gpt-4-1/

[25] huggingface.co/google/metricx-24-hybrid-xl-v2p6

[26] huggingface.co/Unbabel/XCOMET-XL

[27] huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

references by excluding CometKiwi-XL from the AUTORANK computation. This avoids redundancy, as the other hybrid metrics (MetricX-24-Hybrid-XL and XCOMET-XL) can also run in a reference-less (QE) mode to provide the necessary QE signal.

Low-resource exception. For the two lowest-resource languages in the testset, i.e., Bhojpuri and Maasai, we rely solely on `chrF++` (Popović, 2017), computed with `sacrebleu`[28] (Post, 2018). This approach was chosen because the reliability of our main metrics is unestablished for these languages (Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhujan et al., 2025), whereas human references required for `chrF++`[29] were available. Moreover, our cross-metric correlation study—based on Pearson correlations of paragraph-level scores across all systems within each language pair—shows that Bhojpuri, Maasai, and Marathi have the weakest inter-metric agreement (Kocmi et al., 2025b). This observation further supports our use of `chrF++` for Bhojpuri and Maasai. For Marathi, reference translations are not available, so its evaluation necessarily relies on QE metrics.

From system-level scores to AUTORANK. To combine the metrics into a single score, we first normalize them using median-interpercentile scaling to address differences in scale and reduce the influence of low-performing outliers. We then compute the average using equal weights. Finally, we linearly rescale the results to the range from 1 to N systems. A detailed description is provided below:

Let S be the set of submitted systems for a given language pair, $|S| = N$, and let M be the set of automatic metrics used for that language pair (for Bhojpuri and Maasai, $|M| = 1$). For each metric $m \in M$ and system $s \in S$, we compute a system-level score $x_s^{(m)}$ as the average of that metric over all available test segments. To combine scores across metrics, we first map them to a common scale; however, classical min–max normalization is highly sensitive to outliers. In fact, anchoring the scale at the single worst and best system allows an extremely low-scoring outlier to set the lower bound and compress the remaining scores into a narrow band near the top, obscuring meaningful differences among competitive systems. To down-weight extremes without discarding any system, we apply a median-interpercentile scaling to each metric m:

$$\tilde{x}^{(m)} = \text{median}\left\{x_s^{(m)} \mid s \in S\right\}, \quad (1a)$$

$$D^{(m)} = \max\left(\varepsilon,\, Q_{100}^{(m)} - Q_{25}^{(m)}\right), \quad (1b)$$

$$z_s^{(m)} = \frac{x_s^{(m)} - \tilde{x}^{(m)}}{D^{(m)}}. \quad (1c)$$

Where $\varepsilon > 0$ and $Q_p^{(m)}$ denotes the p-th percentile of $\{x_s^{(m)} : s \in S\}$. Importantly, Eq. (1) is continuous and monotonic: it keeps all systems and preserves their order within each metric. Then, for each system, we average the robust-scaled values across metrics:

$$\bar{z}_s = \frac{1}{|M|} \sum_{m \in M} z_s^{(m)}. \quad (2)$$

Averaging after robust scaling yields a single comparable score that preserves the magnitude of performance differences between systems (in standardized units) while preventing any single metric's outliers from dominating. Finally, for readability and to follow the WMT convention from last year (lower is better in AUTORANK, i.e., 1 is best and N worst), we apply a final linear mapping to the set $\{\bar{z}_s\}_{s \in S}$. Specifically, within $\{\bar{z}_s\}_{s \in S}$ the system with the highest average score is assigned 1, the system with the lowest average score is assigned N, and all remaining systems are placed linearly between these two endpoints. This remapping is applied only once—after the cross-metric aggregation—so it preserves the ordering and relative spacing between systems while retaining the outlier mitigation provided by the robust scaling. We refer to the resulting value as AUTORANK in the various tables.

Selecting systems for human evaluation. Following the procedure established in the preliminary report (Kocmi et al., 2025b), we use AUTORANK to select the subset of systems that undergo manual assessment. The target size is 18 systems per language pair, although this number can be higher in certain cases. Selection proceeds in two steps:

1. **Prioritizing constrained systems.** We first select the top-8 performing constrained systems according to AUTORANK.
2. **Filling to target.** We then add the best remaining systems—constrained or unconstrained in

[28] github.com/mjpost/sacrebleu

[29] SacreBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1`.

order of AUTORANK—until the language pair reaches the target number of systems.

Systems not selected for human evaluation keep their AUTORANK ranking as the official result for that language pair (Kocmi et al., 2025b).

5.1 Identification of Error Cases and Impact on Automatic Metrics

Generating text in wrong language. A problem that is particularly common with LLMs used for MT is generating text in the wrong language, either because of copying the source or getting the wrong target language (Bawden and Yvon, 2023; Zhang et al., 2023). To diagnose this issue, we ran language identification on the outputs of the systems using `fasttext` (Joulin et al., 2017b,a), considering a system-language pair problematic if the target language is incorrectly detected in more than 10% of examples.

The most problematic target languages detected by `fasttext` were Serbian, both in Latin script (13 systems) and Cyrillic script (9 systems), Kannada (6 systems), and Marathi (6 systems). Many of these outputs were incorrectly detected as closely related languages: Croatian, Serbo-Croatian and Bosnian for Serbian (particularly when in Latin script), and Hindi for Marathi. In the case of Serbian, these misclassifications appear to be mainly a consequence of the language identification tool's bias: when the same outputs are transcribed into Cyrillic, Serbian is correctly detected. For Marathi, which shares a script with Hindi, the issue is more substantive, with outputs containing mixed or predominantly Hindi content. Kannada outputs, written in a different script, are visually telling: many contain Devanagari script (used for Hindi) or a mix of Devanagari and Kannada characters.

Copying the source text. Another commonly observed issue is generating output in the source language, particularly directly copying the source text. In our case, this mostly corresponds to English source texts, which is unsurprising, given that English was by far the most common source language. Aside from clear failure cases (e.g. NLLB and Gemma-3-12B outputs indicating "FAILED"), many of output language errors, particularly from CommandR7B and EuroLLM-9B (other models to a lesser extent), come from copying large portions of the source text. In some cases, this copied content is mixed with text in the correct target language.

Impact of incorrect language and source copying on automatic evaluation. Copying source text instead of translating it can pose challenges for automatic evaluation tools, particularly those that reward semantic similarity without taking into account the intended target language. In AUTORANK, this applies to the CometKiwi-XL metric, which calculates scores solely based on the source text and does not account for the correctness of the target language. To estimate the impact of source copying on this QE metric, we run a controlled experiment comparing predicted scores for (i) reference texts, (ii) source texts and (iii) different degrees of mixing between source and reference texts to simulate different levels of copying.[30] The results in Table 7 show that, in this setup, source copying does not artificially inflate scores as might be expected. Scores for source-only outputs are well below those for the reference, and the greater the proportion of copied source text in the reference, the lower the score. Interestingly, partial copying leads to approximately the same score degradation as only partial translation, i.e., when the same copied portion is empty.

The metrics generally show positive correlation with each other. However, for a significant number of paragraphs, metrics diverge in their estimation of translation quality. To better understand whether specific patterns exist related to source copying or incorrect language generation, we sought to find instances where metric scores diverged for individual paragraphs. For each paragraph (specific to a language direction), we ranked the system scores and assigned each to a decile (1–10) based on their relative performance. Repeating this process for each metric, we defined the spread for a given paragraph as the difference between the highest and lowest decile across metrics. We observed a large number of divergent instances, even with wide spreads (see Table 8), indicating that metrics behave very differently in certain scenarios. While some metrics are clearly more similar to each other (e.g., LLM-based metrics and the two COMET-based metrics), the dissimilarities in rankings are present between all metrics.

Refer to Figure 2 for how often each metric appears as the score in the lowest or highest decile

[30] For each paragraph we took the first 25%, 50% or 75% of whitespace separated tokens from the source text and concatenated them with the last 75%, 50% or 25% of the reference text. We also tested using only the initial parts of the source text without concatenating the reference texts.

Hypothesis	Score
Source	0.23
Ref.	0.55
Source (⊢25%)	0.27
Source (⊢50%)	0.26
Source (⊢75%)	0.26
Source (⊢25%) · Ref. (⊣75%)	0.46
Source (⊢50%) · Ref. (⊣50%)	0.38
Source (⊢75%) · Ref. (⊣25%)	0.32
Ref. (⊢25%)	0.32
Ref. (⊢50%)	0.38
Ref. (⊢75%)	0.45
Ref. (⊣25%)	0.33
Ref. (⊣50%)	0.41
Ref. (⊣75%)	0.46

Table 7: CometKiwi-XL scores for different hypothesis (compared against the source). · indicates concatenation, and percentages indicate a percentage of the total text tokens taken either consecutively from the start of the text (⊢) or the end (⊣).

relative to others when spreads are greater than 5. No consistent pattern emerges: some metrics appear to reward source copying in certain examples, but not in others, and the overall correlation with the amount of copying and the metric scores is often small. For example, CometKiwi-XL is negatively correlated with the amount of copying (as measured by the 4-gram precision count calculated by `sBLEU`) and XCOMET-XL shows the strongest positive correlation at 0.061 (Pearson coefficient). Similarly, correlations between scores and correct target language decision are also low.

Some further investigation is necessary in following years to better understand the impact of errors on the different metrics, particularly if they are to be used for automatic ranking. Notably, preliminary experiments indicate that the different metrics used are not necessarily well aligned in terms of scores of individual paragraphs, and a more in-depth study could help to understand discrepancies.

	0	2	4	6	8
#paragraphs	370k	320k	206k	89k	22k
%paragraphs	100	86.35	55.73	23.97	5.83

Table 8: Number of paragraphs for which the decile spread across metrics is equal to or higher than thresholds 0–9.

	CometKiwi-XL	GEMBA-ESA -CMDA	GEMBA-ESA -GPT4.1	MetricX-24 -Hybrid-XL	XCOMET-XL
CometKiwi-XL	0	7.1k	6.1k	5.7k	7.3k
GEMBA-ESA-CMDA	9.1k	0	4.9k	9.4k	12.4k
GEMBA-ESA-GPT4.1	7.0k	4.7k	0	7.0k	9.6k
MetricX-24-Hybrid-XL	5.0k	7.0k	5.2k	0	7.0k
XCOMET-XL	3.7k	9.7k	7.6k	5.3k	0

Figure 2: Disagreement matrix showing the number of paragraph–system instances where metric i assigned the minimum decile and metric j assigned the maximum decile. Only cases where the spread between the highest and lowest decile across metrics was greater than 5 are included.

6 Human Evaluation

The human evaluation is done primarily using Error Span Annotation (ESA; Kocmi et al., 2024b). For English→Korean and Japanese→Chinese we rely on the Multidimensional Quality Metrics (MQM; Lommel et al., 2014a).

The ESA Protocol. The annotators (professional translators but not experts in MQM/ESA-style annotations) were asked to mark each error as well as its severity, "Minor" or "Major". In addition, the annotators were also asked to assign a score from 0 to 100, similar to Direct Assessment (DA), to the whole annotation segments (usually a paragraph). However, the ESA score should be more robust than DA alone because the annotators are primed by the highlighted errors at the time of the scoring.

The ESA interface. The interface is shown in Figure 1 with annotator instructions in Appendix A. At the start of annotation, each annotator was exposed to an interactive tutorial where they were asked to interact with the system. The source for the speech domain is a video which is shown in a native HTML video player. The output of the ESA annotation is a list of errors and their severity (minor or major) and the final score from 0 to 100 for each segment.

Task setup. The whole annotation was split into "tasks" where each task had a balanced number of words to make it approximately 1 hour long. Each task is done by a single annotator and contains segments from a single domain but contains output from multiple systems. In contrast to previous years, we do not include a quality control check due

to their annotation costs and low reliability (Zouhar et al., 2025a). Instead, we include "control tasks" for each language, which is the same task that each participating annotator has to fill out exactly once. Because these control tasks are fixed, this allows us to model annotator bias and reliability. Finally, each segment is annotated exactly twice, which can be used to estimate inter-annotator agreement and is especially useful for the metrics shared task (Lavie et al., 2025). See list of changes in contrast to previous version in Appendix A.

Diversity sampling. From the whole testset which all systems translated, we select a subset to human-evaluate. Specifically, we select a 50% of the original data which contains sources that lead to the most diverse translations (as measured by average pairwise ChrF). This ensures that we do not spend the evaluation budget on segments that have very similar translations, which contribute less to the final system ranking (Zouhar et al., 2025b).

The MQM protocol. MQM (Multidimensional Quality Metrics; Lommel et al. (2014b); Freitag et al. (2021)) is the translation evaluation framework that ESA is based on. Professional translators annotate error spans, assigning to each a severity (Major or Minor) and a category from a two-level error hierarchy (e.g. Accuracy/Mistranslation or Fluency/Grammar). Instead of then asking annotators to assign a numeric score to each segment's translation, scores are automatically calculated by applying a severity- and category-dependent weighting scheme to each error and summing the results.

MQM interface. MQM ratings were collected using the open-source Anthea[31] framework. Similarly to the ESA annotations, for the speech domain the video was used as the source side of the evaluation.

Task setup. Due to concerns with rater fatigue, steps were taken to limit the expected time for any individual MQM rating task. To this end, documents were truncated at paragraph boundaries to include no more than 12 source sentences; if the first paragraph contained more than 12 sentences, the document was skipped. For the literary domain in particular, to avoid truncating the vast majority of segments in the very-long documents, the text was instead split into chunks of one or more paragraphs up to the 12-sentence limit, and each chunk was treated as a separate document for the purposes of conducting the evaluation.

[31] github.com/google-research/google-research/tree/master/anthea

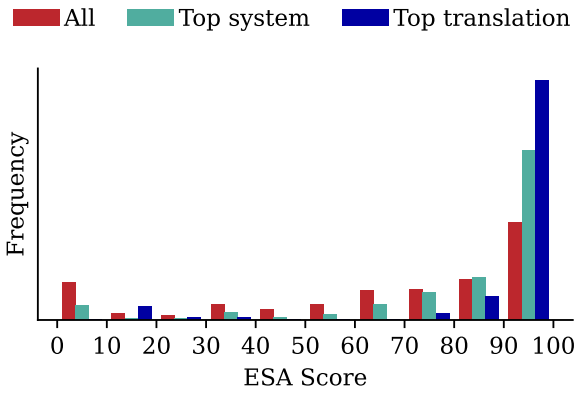


Figure 3: Distribution of final human segment-level scores (ESA) of all systems, top-system in each language, and top translation for each segment. The buckets have a width of 10 (corresponding to xticks). Results include all languages.

6.1 Human Evaluation Analysis

Score distribution. We analyze the score distribution from three perspectives: (1) scores across all systems, (2) scores of the top system in each language pair, and (3) score of top translation for each source segment. While Figure 3 shows that a near-perfect translation (a score of 90-100) is achievable for most source segments by at least one system, the performance of any single system is more modest. Even the top-performing system often fails to achieve a score above 90, a trend that is naturally more pronounced for lower-ranked systems.

Cost and reliability. We also analyze the annotation process itself, focusing on the trade-off between cost and reliability (Kocmi et al., 2024b; Zouhar et al., 2025a). Table 9 provides an overview of annotation volume, time (as a proxy for cost), and inter-annotator agreement (pairwise Pearson). Our analysis finds that while annotation speed and the number of annotated errors vary considerably, neither is predictive of inter-annotator agreement.

Domain and language difficulty. In Table 10, we present the score of the top-performing system for each language and domain. Focusing on the top-performing system mitigates the effect of low-performing outliers. While absolute scores are not directly comparable across languages (due to dif-

	Annotations	Ann./Sys.	Time/Seg.	Time/Word	Minor/Major	Annotators	IAA
English→Czech	8480	424	77.2s	0.9s	2.2/0.9	17	0.41
English→Masai	7030	370	56.9s	0.7s	0.2/1.2	4	0.17
English→Serbian (Cyrilic)	7828	412	88.5s	0.9s	2.0/1.6	4	0.57
English→Japanese	7182	378	89.2s	1.0s	1.0/0.2	40	0.28
English→Ukrainian	7562	398	23.0s	0.3s	0.8/0.3	2	0.64
English→Estonian	7220	380	90.1s	1.0s	2.7/1.1	8	0.53
English→Arabic (Egyptian)	7562	398	40.3s	0.5s	1.3/0.4	3	0.96
Czech→Ukrainian	8550	450	36.1s	0.5s	0.9/0.7	7	0.35
Czech→German	9702	462	81.2s	1.1s	2.0/1.8	12	0.52
English→Russian	7600	400	68.3s	1.5s	1.8/0.9	6	0.36
English→Bhojpuri	7182	378	96.8s	1.3s	1.7/2.0	4	0.85
English→Italian	7740	430	100.7s	1.0s	1.4/0.6	7	0.51
English→Chinese	7524	396	96.2s	1.5s	1.6/0.4	39	0.23
English→Icelandic	7676	404	62.0s	0.7s	3.2/1.7	5	0.69

Table 9: Overview of collected human evaluation data: Number of annotations, Number of annotations per system, Time per segment, Time per source word, Minor and major errors per segment, Number of annotators, Inter-annotator agreement (pairwise Pearson correlation on control subset that is annotated by all).

	Literary	Speech	Social	News	Avg.
En.→Czech	96.1	87.7	86.5	89.9	90.1
En.→Maasai	17.5	8.7	10.9	8.8	11.5
En.→Ukrainian	95.2	87.2	90.2	91.2	90.9
En.→Bhojpuri	98.8	88.3	95.1	95.9	94.5
En.→Japanese	98.4	86.3	92.1	88.9	91.4
En.→Icelandic	87.4	87.3	86.8	88.4	87.5
En.→Estonian	96.8	71.0	88.9	83.0	84.9
En.→Chinese	98.3	83.7	92.4	87.7	90.5
En.→Russian	94.8	77.0	83.7	83.6	84.8
En.→Arabic (Egy.)	88.4	80.8	84.3	74.7	82.0
En.→Serbian (Cyr.)	98.6	89.5	96.0	93.5	94.4
En.→Italian	87.0	71.9	83.4	86.1	82.1
Cz.→Ukrainian		89.4	92.9	94.7	92.3
Cz.→German		90.4	95.7	88.8	91.6

Table 10: Human evaluation scores for the top-performing system per language, by domain.

ferent sets of systems, annotators, and sources), we again observe a consistent pattern (Kocmi et al., 2024a): the speech domain receives the lowest scores, suggesting it is the most challenging, likely due to its reliance on ASR-generated text. The social and news domains follow in difficulty. Surprisingly, the literary domain achieved the highest scores.[32]

Most language pairs show similar difficulty patterns, with English→Maasai as a notable outlier. Overall, English→Arabic (Egyptian) and English→Italian proved to be the most challenging.

[32]This could be due to two factors: first, its source texts were not subject to difficulty sampling due to limited number of stories. Second, an evaluation that asks annotators to mark errors may award high scores to translations that are technically error-free, even if they do not fully capture stylistic qualities such as the author's voice or the reader's enjoyment (Carpuat et al., 2025).

7 Official Ranking Results

We now describe how we compute the final ranking and then discuss the final results and potential issues. The ranking is presented in tabular form in Section 7.4.

7.1 Human Ranking Computation

We calculate three different scores: the human ESA or MQM score, rank, and the cluster. The human ESA or MQM scores are the micro-average of the segment-level scores. This disregards any domain balancing, though we show per-domain results in Appendix D. For the statistical analysis and clustering, we use the Wilcoxon signed-rank test, a paired non-parametric test (Wilcoxon, 1945), with $p < 0.05$. The rank ranges differ from last year's implementation. Systems are sorted by their average human score and for a system in row i we define its rank range $\langle i_\downarrow, i_\uparrow \rangle$ as follows: $i_\downarrow := \max\{j | j < i, \text{significant(i, j)}\} + 1$ and $i_\uparrow := \min\{j | j > i, \text{significant(j, i)}\} - 1$. In words, the ranks expand from i until a system that is statistically distinguishable is encountered. Lastly, the clusters are the maximal partition of systems such that ranks of systems from one cluster do not overlap with ranks of systems in another cluster.

7.2 Human Evaluation Discussion

In this section, we discuss the results of the human evaluation presented in Section 7.3 (constrained systems only) and in Section 7.4 (all systems).

7.3 Ranking of Constrained Systems

English→Chinese

Rank	System	Human
1-1	Algharb	88.4
2-2	Shy-hunyuan-MT	88.2
3-3	**Human**	82.1
4-5	SRPOL ⃠	77.7
4-6	IRB-MT	76.5
5-6	RuZh ?	75.7
7-7	Laniqo	70.5

English→Ukrainian

Rank	System	Human
1-1	Algharb	90.0
2-2	Shy-hunyuan-MT	88.4
3-3	**Human**	87.3
4-4	TowerPlus-9B[M]	84.2
5-5	IRB-MT	82.9
6-7	SRPOL	79.9
6-7	Laniqo	79.8

English→Italian

Rank	System	Human
1-1	Shy-hunyuan-MT	78.7
2-3	TowerPlus-9B[M]	61.2
2-3	IRB-MT	60.3
4-6	SalamandraTA	57.5
4-6	AyaExpanse-8B	57.0
4-6	EuroLLM-9B[M]	56.6
7-8	Gemma-3-12B	53.6
7-8	Laniqo	53.4

English→Czech

Rank	System	Human
1-1	Shy-hunyuan-MT	87.1
2-2	**Human**	84.5
3-3	Algharb	76.7
4-4	CUNI-MH-v2	71.0
5-6	SRPOL	67.5
5-7	Laniqo	66.1
6-7	TowerPlus-9B[M]	65.8
8-8	SalamandraTA	60.3

English→Arabic (Egyptian)

Rank	System	Human
1-1	**Human**	78.5
2-2	IRB-MT	51.9
3-5	CommandR7B	3.7
3-6	Algharb	3.2
3-6	Shy-hunyuan-MT	3.2
4-6	AyaExpanse-8B	2.0
7-7	SRPOL ⃠	0.9

English→Russian

Rank	System	Human
1-1	Shy-hunyuan-MT	80.2
2-2	Algharb	73.3
3-3	**Human**	70.5
4-4	IRB-MT	65.4
5-7	RuZh ?	57.9
5-7	SRPOL	56.9
5-7	Laniqo	56.2

English→Estonian

Rank	System	Human
1-1	**Human**	83.1
2-3	Algharb	70.4
2-3	Shy-hunyuan-MT	70.3
4-5	SRPOL	49.4
4-6	Laniqo	48.6
5-6	SalamandraTA	46.7
7-7	IRB-MT	32.4

English→Bhojpuri

Rank	System	Human
1-2	**Human**	92.6
1-2	Algharb	91.1
3-3	NLLB	75.6
4-4	COILD-BHO	68.7
5-5	IRB-MT	59.6
6-6	SalamandraTA	35.7
7-7	Shy-hunyuan-MT	1.7

Czech→German

Rank	System	Human
1-1	Shy-hunyuan-MT	87.2
2-2	**Human**	82.8
3-4	Algharb	80.9
3-4	TowerPlus-9B[M]	79.8
5-7	CUNI-MH-v2	77.2
5-7	Gemma-3-12B	76.8
5-7	SRPOL	76.7
8-9	IRB-MT	71.7
8-9	Laniqo	70.0

Czech→Ukrainian

Rank	System	Human
1-1	Shy-hunyuan-MT	91.8
2-2	**Human**	90.1
3-4	TowerPlus-9B[M]	85.3
3-5	Algharb	84.1
4-6	Laniqo	83.4
5-6	IRB-MT	82.7
7-7	SRPOL	80.8

Japanese→Chinese

Rank	System	Human
1-1	**Human**	-3.5
2-3	Algharb	-5.8
2-3	Shy-hunyuan-MT	-6.1
4-5	NTTSU	-11.3
4-5	TowerPlus-9B[M]	-13.3
6-6	IRB-MT	-13.9
7-7	Laniqo	-18.3

English→Serbian (Cyrilic)

Rank	System	Human
1-1	Shy-hunyuan-MT	92.2
2-2	**Human**	88.7
3-3	IRB-MT	77.6
4-5	SalamandraTA	75.5
4-5	Gemma-3-12B	74.8
6-7	CUNI-SFT	60.9
6-7	Llama-3.1-8B ⃠	58.4
8-8	NLLB	53.5
9-9	EuroLLM-9B[M] ⃠	41.8

English→Icelandic

Rank	System	Human
1-1	**Human**	87.5
2-2	Shy-hunyuan-MT	63.2
3-3	TowerPlus-9B[M]	57.4
4-4	AMI	39.9
5-5	SalamandraTA	31.3
6-6	NLLB	24.1
7-7	IRB-MT	20.7
8-8	Gemma-3-12B	16.5
9-9	Llama-3.1-8B ⃠	10.5

English→Korean

Rank	System	Human
1-2	**Human**	-1.9
1-2	Shy-hunyuan-MT	-2.5
3-4	Algharb	-4.4
3-5	IRB-MT	-5.6
4-5	Gemma-3-12B	-5.9
6-6	TowerPlus-9B[M]	-7.2
7-7	Laniqo	-9.1

English→Japanese

Rank	System	Human
1-1	**Human**	89.2
2-2	Algharb	85.7
3-3	Shy-hunyuan-MT	79.9
4-5	KIKIS	76.2
4-5	Systran	75.6
6-6	NTTSU	73.3
7-7	Laniqo	67.8

English→Masai

Rank	System	Human
1-1	**Human**	9.6
2-3	AyaExpanse-8B ⃠	6.0
2-3	Shy-hunyuan-MT ⃠	4.8
4-6	Llama-3.1-8B ⃠	3.0
4-6	Gemma-3-12B ?	3.0
4-6	Qwen2.5-7B ?	2.8
7-9	CommandR7B ⃠	1.6
7-9	TowerPlus-9B[M] ⃠	0.8
7-9	EuroLLM-9B[M] ⃠	0.7

Human evaluation shifts overall ranking. Overall, the best-performing model in the human evaluation is Gemini 2.5 Pro (see Figure 5).[33] It places in the top cluster for 14 of the 16 evaluated language pairs and is on par with or surpasses human translation in 10 of those pairs.[34]

This result differs notably from the AUTORANK results (Kocmi et al., 2025b), where Shy-hunyan-MT ranked first for all but one language pair (English→Bhojpuri). The discrepancy may be due to Shy-hunyan-MT's use of GRPO with XCOMET-XXL and GEMBA with DeepSeek V3 as its training signals.[35] However, despite its lower ranking in the human evaluation, Shy-hunyan-MT remains the best-performing *constrained system*, winning in this category for 11 language pairs. The second best constrained system is Algharb, which ranks first in six language pairs.

Finally, human translations, often treated as a "gold standard," place in the top cluster for only six of the 15 language pairs where they were available. This finding highlights the inherent difficulty of translation, though it could also reflect the stylistic or lexical preferences of the annotators.

7.4 Official Ranking Results Tables

Results tables legend

The human score is the micro-average of human judgments across all domains and double annotations (single annotations for MQM language pairs). AutoRank is calculated from automatic metrics as per (Kocmi et al., 2025b). Significance testing is done using a Wilcoxon signed rank test with a p-value threshold of 5%. The rank range for the ith model begins as $\langle i, i \rangle$ and is expanded in both directions until a significant difference is found. Clusters are formed such that their constituent rank ranges do not overlap.

Systems are either constrained (white), or unconstrained (gray). Systems that do not officially support the language pair are marked with 🚫 and those where language support cannot be verified are marked with ❓. The [M] suffix marks systems (submitted by the WMT organizers) that were trained/-tuned with specific MT instructions, but prompted without these specific instructions (using a generic setup, same for all LLMs, see Section 4.2), which could disadvantage these systems. See Appendix D for a per-domain breakdown of system performances.

English→Arabic (Egyptian)

Rank	System	Human	AutoRank
1-1	**Human**	78.5	
2-2	GPT-4.1	77.0	6.7
3-3	CommandA	74.0	8.6
4-4	Gemini-2.5-Pro	60.6	5.8
5-6	DeepSeek-V3❓	56.8	7.0
5-6	Claude-4	55.7	7.8
7-7	IRB-MT	51.9	11.1
8-9	Mistral-Medium	36.0	7.7
8-9	CommandA-WMT	34.6	4.1
10-10	UvA-MT	29.0	4.2
11-14	CommandR7B	3.7	11.6
11-14	GemTrans	3.7	3.5
11-16	Algharb	3.2	2.7
11-16	Shy-hunyuan-MT	3.2	1.0
13-16	AyaExpanse-8B	2.0	9.9
12-16	ONLINE-B	1.7	6.5
17-19	Yolu	1.4	5.5
15-18	Wenyiil	1.4	2.5
19-19	SRPOL🚫	0.9	8.1
20-39	19 systems not human-evaluated		...

English→Estonian

Rank	System	Human	AutoRank
1-1	**Human**	83.1	
2-2	Gemini-2.5-Pro	78.8	2.5
3-4	Wenyiil	72.6	2.6
3-4	GPT-4.1	72.2	3.0
5-6	Algharb	70.4	3.9
5-6	Shy-hunyuan-MT	70.3	1.0
7-8	ONLINE-B	60.2	6.0
7-8	Yolu	59.5	3.8
9-9	TranssionTranslate❓	57.1	7.3
10-11	Claude-4❓	53.0	6.5
10-12	GemTrans	51.7	5.1
11-14	CommandA-WMT🚫	50.1	6.1
12-15	SRPOL	49.4	5.7
12-17	Laniqo	48.6	5.2
13-17	EuroLLM-22B-pre.[M]	47.2	8.1
14-18	SalamandraTA	46.7	6.3
14-18	UvA-MT	46.4	5.9
16-18	Gemma-3-27B	45.9	7.6
19-19	IRB-MT	32.4	11.4
20-40	20 systems not human-evaluated		...

English→Bhojpuri

Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	94.9	1.0
2-3	**Human**	92.6	
2-3	Algharb	91.1	2.8
4-4	Wenyiil	90.9	2.5
5-6	Claude-4❓	83.2	4.5
5-6	GPT-4.1❓	82.8	5.5
7-8	TranssionTranslate❓	79.5	4.3
7-10	DeepSeek-V3❓	77.3	5.1
8-10	Llama-4-Maverick	76.4	6.5
8-10	NLLB	75.6	6.6
11-12	CommandA🚫	72.6	6.5
11-12	Yolu	72.4	5.7
13-14	TranssionMT	70.1	6.2
13-15	COILD-BHO	68.7	8.9
14-15	ONLINE-B	67.2	4.1
16-16	IRB-MT	59.6	11.4
17-17	Gemma-3-27B❓	56.0	8.3
18-18	SalamandraTA	35.7	12.1
19-19	Shy-hunyuan-MT	1.7	11.5
20-37	17 systems not human-evaluated		...

[33] The translations with Gemini 2.5 Pro were collected with the "thinking" mode enabled, and it is unclear how much this contributed to its overall performance.

[34] Note that while human translations for WMT are prepared by professional translators, they are not necessarily free of errors or undergo the same level of post-editing as human translations used in high-stake scenarios (e.g., sworn translation) or published translations (e.g., literary translation).

[35] Similar shifts, i.e., a higher automatic rank and lower placement after human evaluation, are visible for other systems, some of which also use metrics as a signal in training or reranking (see Figure 4).

English→Masai

Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro❓	9.8	6.1
2-2	**Human**	9.6	
3-3	Claude-4❓	7.7	2.6
4-6	AyaExpanse-8B🚫	6.0	8.2
4-5	Llama-4-Maverick🚫	5.2	3.2
6-6	Shy-hunyuan-MT🚫	4.8	1.0
7-13	AyaExpanse-32B🚫	3.1	7.1
4-8	DeepSeek-V3❓	3.0	6.2
9-13	Llama-3.1-8B🚫	3.0	8.1
9-13	Gemma-3-12B❓	3.0	8.8
9-13	Qwen2.5-7B❓	2.8	8.6
9-13	Qwen3-235B🚫	2.7	3.0
9-13	TranssionMT	2.5	5.9
14-18	CommandR7B🚫	1.6	4.3
14-18	CommandA-WMT🚫	1.5	6.4
14-16	CommandA🚫	1.3	7.9
17-18	TowerPlus-9B[M]🚫	0.8	5.3
17-18	EuroLLM-9B[M]🚫	0.7	8.2
19-19	EuroLLM-22B-pre.[M]🚫	0.5	8.2
20-29	9 systems not human-evaluated		...

English→Russian

Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	83.4	4.4
2-2	Shy-hunyuan-MT	80.2	1.0
3-5	Wenyiil	78.2	4.8
3-5	GPT-4.1	76.2	5.4
3-5	Claude-4	75.9	8.7
6-9	DeepSeek-V3❓	73.6	5.7
5-8	Algharb	73.3	5.2
6-9	CommandA-WMT	73.2	4.2
8-10	Yandex	72.0	4.5
9-11	**Human**	70.5	
10-12	UvA-MT	69.1	4.5
11-14	Qwen3-235B	67.6	8.8
12-15	IRB-MT	65.4	10.1
12-15	Yolu	64.5	6.9
13-16	GemTrans	62.5	5.1
15-16	Gemma-3-27B	61.7	8.9
17-19	RuZh❓	57.9	9.6
17-19	SRPOL	56.9	10.6
17-19	Laniqo	56.2	8.8
20-42	22 systems not human-evaluated		...

English→Ukrainian

Rank	System	Human	AutoRank
1-3	Gemini-2.5-Pro	90.3	3.3
1-3	Algharb	90.0	4.2
1-3	Wenyiil	89.5	3.5
4-5	Shy-hunyuan-MT	88.4	1.0
4-5	GemTrans	88.2	4.6
6-7	GPT-4.1	87.9	3.5
5-8	**Human**	87.3	
7-9	UvA-MT	86.4	4.4
8-13	CommandA-WMT	86.3	3.9
9-13	Llama-4-Maverick	86.2	8.8
9-13	DeepSeek-V3❓	85.8	5.0
9-14	Claude-4❓	85.6	7.0
9-13	Yolu	85.4	6.0
14-16	Mistral-Medium❓	84.5	6.0
14-16	TowerPlus-9B[M]	84.2	8.8
14-16	CommandA	84.0	7.4
17-17	IRB-MT	82.9	8.2
18-19	SRPOL	79.9	8.4
18-19	Laniqo	79.8	7.7
20-44	24 systems not human-evaluated		...

English→Italian

Rank	System	Human	AutoRank
1-4	Gemini-2.5-Pro	79.4	4.4
1-4	GemTrans	79.4	5.2
1-4	GPT-4.1	79.0	4.5
1-4	Shy-hunyuan-MT	78.7	1.0
5-7	CommandA-WMT	75.5	2.6
5-8	Mistral-Medium❓	73.8	7.1
5-10	CommandA	73.2	8.4
6-10	Claude-4	72.1	8.4
7-10	UvA-MT	71.8	5.3
7-10	DeepSeek-V3❓	71.7	6.1
11-11	Qwen3-235B	67.0	7.2
12-13	TowerPlus-9B[M]	61.2	11.3
12-13	IRB-MT	60.3	10.2
14-16	SalamandraTA	57.5	10.3
14-16	AyaExpanse-8B	57.0	14.9
14-16	EuroLLM-9B[M]	56.6	15.2
17-18	Gemma-3-12B	53.6	15.5
17-18	Laniqo	53.4	7.6
19-34	15 systems not human-evaluated		...

English→Icelandic

Rank	System	Human	AutoRank
1-1	**Human**	87.5	
2-2	Gemini-2.5-Pro	77.6	1.8
3-4	Erlendur	68.3	2.2
3-4	GPT-4.1	68.0	1.9
5-5	Shy-hunyuan-MT	63.2	1.0
6-6	TowerPlus-9B[M]	57.4	3.9
7-7	ONLINE-B	51.8	4.4
8-10	Claude-4❓	47.8	5.2
8-10	TowerPlus-72B[M]	46.3	5.7
8-10	TranssionTranslate❓	46.2	5.8
11-11	AMI	39.9	7.4
12-12	GemTrans	34.8	7.0
13-14	SalamandraTA	31.3	8.6
13-15	UvA-MT	30.6	6.8
14-15	CommandA-WMT🚫	29.0	6.8
16-16	NLLB	24.1	15.2
17-17	IRB-MT	20.7	11.9
18-18	Gemma-3-12B	16.5	13.8
19-19	Llama-3.1-8B🚫	10.5	24.9
20-35	15 systems not human-evaluated		...

English→Serbian (Cyrilic)

Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	94.2	3.0
2-3	GPT-4.1	92.5	3.4
2-4	Shy-hunyuan-MT	92.2	1.0
3-4	ONLINE-B	90.6	6.1
5-5	Claude-4❓	90.0	6.8
6-6	**Human**	88.7	
7-7	TranssionTranslate❓	85.1	8.0
8-9	GemTrans	81.5	4.6
8-9	DeepSeek-V3❓	78.7	8.6
10-11	IRB-MT	77.6	9.9
10-15	DLUT_GTCOM	77.2	9.3
11-14	CommandA-WMT🚫	76.5	7.0
10-15	UvA-MT	76.2	5.8
11-15	SalamandraTA	75.5	8.8
13-15	Gemma-3-12B	74.8	12.1
16-17	CUNI-SFT	60.9	13.5
16-17	Llama-3.1-8B🚫	58.4	19.4
18-18	NLLB	53.5	19.8
19-19	EuroLLM-9B[M]🚫	41.8	22.3
20-34	14 systems not human-evaluated		...

Czech→German

Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	90.7	2.5
2-4	GPT-4.1	89.5	2.4
2-4	Claude-4	88.8	4.8
2-6	DeepSeek-V3❓	88.1	3.5
4-7	Shy-hunyuan-MT	87.2	1.0
4-8	Mistral-Medium	87.0	4.2
5-7	CommandA	86.8	4.8
8-8	CommandA-WMT	85.6	2.1
9-12	**Human**	82.8	
9-13	GemTrans	82.6	6.3
9-13	Gemma-3-27B	82.0	7.2
9-13	Wenyiil	82.0	10.9
10-15	Algharb	80.9	13.2
13-15	TowerPlus-9B[M]	79.8	10.3
13-15	UvA-MT	79.5	7.0
16-19	CUNI-MH-v2	77.2	14.2
16-18	Gemma-3-12B	76.8	11.5
16-18	SRPOL	76.7	11.0
19-19	Yolu	75.3	9.3
20-21	IRB-MT	71.7	12.4
20-21	Laniqo	70.0	10.3
22-42	20 systems not human-evaluated		...

English→Czech

Rank	System	Human	AutoRank
1-1	Gemini-2.5-Pro	88.7	3.4
2-2	Shy-hunyuan-MT	87.1	1.0
3-4	DeepSeek-V3❓	85.1	5.1
3-4	**Human**	84.5	
5-6	CommandA-WMT	82.6	3.6
5-6	Wenyiil	82.4	4.5
7-9	GPT-4.1	80.8	4.0
7-9	Mistral-Medium❓	80.4	7.1
7-10	Claude-4❓	79.6	9.0
9-11	UvA-MT	78.6	6.5
10-14	Algharb	76.7	6.4
11-14	CommandA	76.4	8.8
11-15	Yolu	75.6	6.3
11-15	Gemma-3-27B	75.6	9.2
13-15	GemTrans	73.2	5.1
16-16	CUNI-MH-v2	71.0	12.1
17-18	SRPOL	67.5	8.7
17-19	Laniqo	66.1	8.8
18-19	TowerPlus-9B[M]	65.8	11.6
20-20	SalamandraTA	60.3	10.5
21-44	23 systems not human-evaluated		...

English→Chinese

Rank	System	Human	AutoRank
1-1	Algharb	88.4	4.2
2-4	Shy-hunyuan-MT	88.2	1.0
2-5	Claude-4	86.9	7.2
2-5	Wenyiil	86.3	4.0
3-6	DeepSeek-V3	85.0	7.3
5-10	GemTrans	84.4	5.0
6-11	Qwen3-235B	84.0	4.9
5-10	GPT-4.1	84.0	4.7
6-11	Gemini-2.5-Pro	83.8	4.0
5-10	UvA-MT	83.4	6.4
11-13	**Human**	82.1	
11-15	CommandA-WMT	81.3	5.7
11-15	Llama-4-Maverick	80.7	8.1
12-16	Mistral-Medium❓	79.9	5.0
12-16	Yolu	79.0	4.9
14-17	SRPOL🚫	77.7	10.5
16-18	IRB-MT	76.5	9.5
17-18	RuZh❓	75.7	10.6
19-19	Laniqo	70.5	9.3
20-40	20 systems not human-evaluated		...

English→Japanese

Rank	System	Human	AutoRank
1-1	**Human**	89.2	
2-4	Gemini-2.5-Pro	85.8	2.5
2-6	Algharb	85.7	3.3
2-5	Mistral-Medium❓	84.8	5.5
3-6	Wenyiil	84.4	3.0
5-6	GPT-4.1	83.7	2.9
7-7	CommandA-WMT	82.2	3.7
8-12	Shy-hunyuan-MT	79.9	1.0
8-13	DeepSeek-V3❓	79.3	4.7
8-13	Claude-4	79.3	5.8
8-13	UvA-MT	79.3	6.5
8-14	ONLINE-B	78.0	6.3
9-16	In2x❓	77.8	2.3
12-16	GemTrans	76.2	5.6
13-16	KIKIS	76.2	3.2
13-16	Systran	75.6	7.5
17-18	NTTSU	73.3	8.1
17-18	Yolu	72.6	6.1
19-19	Laniqo	67.8	9.5
20-45	25 systems not human-evaluated		...

Czech→Ukrainian

Rank	System	Human	AutoRank
1-2	Gemini-2.5-Pro	92.9	1.1
1-3	GPT-4.1	92.1	1.3
2-3	Shy-hunyuan-MT	91.8	1.0
4-8	GemTrans	90.2	4.4
4-6	**Human**	90.1	
4-10	Mistral-Medium❓	89.4	4.2
6-10	Claude-4❓	89.1	3.7
4-10	DeepSeek-V3❓	89.0	3.2
6-10	CommandA-WMT	88.7	1.3
6-10	Gemma-3-27B	88.6	5.0
11-12	CommandA	86.4	4.6
11-13	Wenyiil	85.7	5.4
12-15	TowerPlus-9B[M]	85.3	7.9
13-16	Algharb	84.1	7.2
13-17	UvA-MT	83.5	5.1
14-17	Laniqo	83.4	7.7
15-17	IRB-MT	82.7	9.1
18-19	SRPOL	80.8	7.8
18-19	Yolu	80.1	6.0
20-44	24 systems not human-evaluated		...

Japanese→Chinese

Rank	System	Human	AutoRank
1-1	**Human**	-3.5	
2-2	Gemini-2.5-Pro	-4.4	3.3
3-6	Algharb	-5.8	4.3
3-7	Claude-4	-5.9	6.4
3-7	Shy-hunyuan-MT	-6.1	1.0
3-7	GPT-4.1	-6.2	4.5
4-7	Wenyiil	-6.9	4.5
8-10	CommandA-WMT	-7.7	5.2
8-10	DeepSeek-V3	-8.1	6.5
8-13	Kaze-MT	-8.6	3.9
10-13	Mistral-Medium🚫	-10.0	6.6
10-13	In2x🚫	-10.0	3.0
10-13	Qwen3-235B	-10.9	7.6
14-15	GemTrans	-10.9	6.6
14-15	NTTSU	-11.3	5.9
16-17	Yolu	-12.6	7.1
16-17	TowerPlus-9B[M]	-13.3	11.5
18-18	IRB-MT	-13.9	12.4
19-19	Laniqo	-18.3	11.3
20-42	22 systems not human-evaluated		...

English→Korean

Rank	System	Human	AutoRank
1-3	**Human**	-1.9	
1-3	Shy-hunyuan-MT	-2.5	1.0
1-3	Gemini-2.5-Pro	-2.7	2.5
4-6	GPT-4.1	-3.3	2.9
4-7	Claude-4	-3.4	4.4
4-7	DeepSeek-V3 🚫	-3.8	5.1
5-10	GemTrans	-4.1	5.0
7-12	CommandA-WMT	-4.3	2.9
5-12	Wenyiil	-4.3	3.0
5-12	Algharb	-4.4	3.1
8-15	Mistral-Medium 🚫	-4.7	6.1
7-15	CommandA	-4.7	6.0
11-16	UvA-MT	-5.2	4.3
11-16	Qwen3-235B	-5.5	6.5
11-16	IRB-MT	-5.6	8.6
13-16	Gemma-3-12B	-5.9	9.2
17-18	TowerPlus-9B[M]	-7.2	10.1
17-18	Yolu	-7.3	7.0
19-19	Laniqo	-9.1	9.2
20-37	17 systems not human-evaluated		...

While best system scores high many stay in the middle. Although the best system almost always achieves a score of 90 or higher, no system (including human references) achieves a perfect score of 100 (see Table 10 and Appendix D). The spread of scores is also quite large: mid-tier models often score 60 or below, and the worst-performing systems can score as low as 0, depending on the language pair and domain (see Figures 6 and 7).

Translation into low-resource languages remains a challenge. The translation quality for Maasai, a low-resource language, is largely unusable. We observe that for this language, systems often produce large portions of their output in Swahili, and the few acceptable spans in Maasai are frequently overlooked by annotators.[36] Additional errors arise from translations that fail to capture the meaning and context of the source text. While human translations into Maasai also score low (9.6 ESA scores), we were able to independently confirm that these are generally understandable by native speakers and often sound natural. However, they may exhibit extensive code-mixing of English and Swahili (Figures 6 and 7). We speculate that our Maasai annotators, when evaluating a large portion of poor-quality system outputs, failed to notice the one translation that was mostly reasonable.

A similar issue occurs with Egyptian Arabic, where systems tend to output Modern Standard Arabic (MSA). This type of error would likely be overlooked by quality estimation metrics, which typically do not incorporate target language labels into their pipeline (though they could potentially be trained to do so).

[36]While many Maasai people speak Swahili, Maasai (or Maa) is a distinct language from a different language family.

7.5 Additional Analysis of English→Serbian translations

In addition to the official human evaluation, an analysis of English→Serbian translations was carried out by an MT researcher with experience in human translation. One part of the analysis deals with the two scripts (Latin and Cyrillic), and another one with translation quality taking into account both errors as well as exceptionally good idiomatically translated parts named "rewards".

Scripts. Due to historical and cultural reasons, the Serbian language is bi-alphabetical, using both Latin and Cyrillic scripts. Serbian Cyrillic alphabet is a highly phonetic alphabet with a one-to-one correspondence between letters and sounds. Serbian Latin alphabet is almost perfectly compatible with Cyrillic with a one-to-one correspondence, except for the three digraphs each representing one sound (Љ ↔ /lj/, Њ ↔ /nj/, Џ ↔ /dž/). Serbian speakers switch easily between the two scripts without much thinking. The choice of the script is partly random but also influenced by the context, medium, or even ideological preferences. However, the scripts should not be mixed within a single text with a few exceptions: URLs or foreign brand names which should be in Latin even in Cyrillic texts. The automatic conversion of Serbian Cyrillic into Latin script is easier than the other way round. The primary reason is the one-to-one character mapping from Cyrillic to Latin, while converting from Latin to Cyrillic introduces ambiguity due to the three previously mentioned digraphs.

Training data for generative NLP including MT are available in both scripts, and it might be challenging to ensure that the outputs are written in the desired script. Therefore, the WMT translations were checked in this aspect, by measuring the percentage of words written in another script.

The results are presented in Table 11. In Cyrillic translations there is always a small percent of words written in Latin, because of URLs, named entities or similar. Overall, there is more mixing in Cyrillic translations: found in more systems and also to the larger extent. The probable reason is that there is more available data in Latin script. Another observation is that some systems use only one script (e.g. ONLINE-B and ONLINE-G only Cyrillic,

% Cyrillic in Latin	
ONLINE-G	98.9
ONLINE-B	98.0
Mistral-7B	12.9
Gemma-3-12B	11.2
TowerPlus-72B	4.9
Gemma-3-27B	3.4
CommandR7B	3.2
IRB-MT	2.8
Qwen3-235B	1.9
Qwen2.5-7B	1.7
TowerPlus-9B	0.4
AyaExpanse-8B	0.3
Yolu	0.0
Wenyiil	0.0
UvA-MT	0.0
TranssionTranslate	0.0
TranssionMT	0.0
Shy	0.0
SalamandraTA	0.0
NLLB	/
Llama-4-Maverick	0.0
Llama-3.1-8B	0.0
IR-MultiagentMT	0.0
GPT-4.1	0.0
GemTrans	0.0
Gemini-2.5-Pro	0.0
EuroLLM-9B	0.0
EuroLLM-22B	0.0
DLUT_GTCOM	/
DeepSeek-V3	0.0
CUNI-SFT	0.0
CommandA	0.0
Claude-4	0.0
AyaExpanse-32B	0.0
Algharb	0.0

(a) % Unexpected Cyrillic script detected in a Latin-script translation.

% Latin in Cyrillic	
CUNI-SFT	100.0
Llama-3.1-8B	99.7
UvA-MT	95.7
AyaExpanse-8B	85.8
CommandR7B	79.2
Qwen2.5-7B	76.9
TowerPlus-9B	69.6
EuroLLM-9B	66.3
EuroLLM-22B	66.0
IRB-MT	25.8
Gemma-3-12B	22.4
Mistral-7B	8.3
TowerPlus-72B	3.6
Shy	2.9
IR-MultiagentMT	2.7
GPT-4.1	2.4
Gemma-3-27B	2.2
ONLINE-B	2.1
Llama-4-Maverick	2.1
DeepSeek-V3	2.1
AyaExpanse-32B	2.0
Gemini-2.5-Pro	1.9
DLUT_GTCOM	1.9
TranssionTranslate	1.7
Claude-4	1.7
CommandA	1.6
Qwen3-235B	1.4
ONLINE-G	1.4
NLLB	1.4
SalamandraTA	1.3
GemTrans	1.1
Yolu	/
Wenyiil	/
TranssionMT	/
Algharb	/

(b) % Unexpected Latin script detected in a Cyrillic-script translation

Table 11: Comparison of script intrusions across Latin and Cyrillic translations.

CUNI-SFT and Llama-3.1-8B only in Latin, UvA-MT almost all in Latin).

We further explored qualitative differences between the outputs in different scripts. The first step was automatic: Cyrillic outputs were converted into Latin and then compared by word bi-gram overlap (F-score). For some systems, there was almost no difference, but for the majority there were notable differences (for example, around 75% overlap score for Gemini and GPT, around 60% for Shy, see the full table can be seen in Appendix Table 20). However, qualitative manual inspection did not identify any major or systematic differences regarding translation quality or types of errors.

Errors and rewards. We next looked into the annotated error spans similarly to Popovic (2021). However, due to discrepancies in error span annotations it was difficult to determine the nature of the annotated spans. For example, a number of non-existing or obviously incorrect words were not marked at all, and overall scores were not lowered at all or only slightly, while there were completely correct passages which are marked as errors. Furthermore, a number of segments seemed to be heavily penalized only for using the Latin script: all words there were marked as errors without looking at actual errors, and the scores were lowered but inconsistently, ranging from 90 to 10. The affected systems were CUNI-SFT, UvA-MT, LLama-3.1-8B and EuroLLM-9B. After qualitative inspection, it seemed that the translation quality of LLama-3.1-8B and EuroLLM-9B was indeed low, while CuNI-SFT and UvA-MT might be underestimated.

For these reasons, a full additional ESA annotation has been carried out on a small selected set: last ten documents from each of the domains, so 40 documents in total. The following nine translations were included: the best ranked systems in the official evaluation (Gemini, GPT, Shy, ONLINE-B, Claude) and the human translation, the two low-ranked systems which were potentially over-penalised for the script (UvA and CUNI), as well as one system from the middle cluster (GemTrans).

For each translated documents, first all errors were marked, and then overall scores were assigned (the same process as in the official evaluation, although no distinction between major and minor errors was made). During the evaluation, the annotator observed that, apart from error spans, there were passages translated exceptionally well, namely idiomatically: diverging from the source, but fully keeping the meaning while sounding completely natural in the target language. These spans were then marked as "rewards".

The results from the additional span annotation are aggregated as word-level error rates and reward rates, namely number of words in the marked spans divided by total number of words (length). Also, a total score is presented, where error spans are subtracted from the length and reward spans are added. In addition, the official scores and error rates are extracted for the 40 selected documents and presented for comparison.

The results can be seen in Table 12. The largest percentage of idiomatic translations can be found in the middle-ranked GemTrans translation (5.67%), followed by human translation (4.17%). Gemini

	official e.		additional evaluation			
	score	%err.	score	%err.	%rew.	total
Gemini	89.3	5.3	87.4	5.82	3.61	97.8
GPT	93.8	2.8	82.8	8.29	0.95	92.6
Shy	86.4	5.7	85.8	7.02	3.52	96.5
ONLINE-B	84.0	7.7	70.7	13.3	0.42	87.1
Claude	85.9	9.1	63.2	15.2	0.29	85.1
Human	86.8	19.2	86.9	7.84	4.19	96.4
GemTrans	84.2	6.6	81.1	9.60	5.67	96.1
UvA	81.7	76.6	64.8	14.7	1.18	86.4
CUNI	54.2	81.2	41.4	27.2	0.18	73.0

Table 12: Results of the additional analysis on the selected set of translations for English→Serbian together with the scores from the official evaluation.

official evaluation		additional evaluation			
score	%err.	score	%err.	%rew.	total
GPT	GPT	Gemini	Gemini	GemTrans	Gemini
Gemini	Gemini	Human	Shy	Human	Shy
Human	Shy	Shy	Human	Gemini	Human
Shy	GemTrans	GPT	GPT	Shy	GemTrans
Claude	ONLINE-B	GemTrans	GemTrans	UvA	GPT
GemTrans	Claude	ONLINE-B	ONLINE-B	GPT	ONLINE-B
ONLINE-B	Human	UvA	UvA	ONLINE-B	UvA
UvA	UvA	Claude	Claude	Claude	Claude
CUNI	CUNI	CUNI	CUNI	CUNI	CUNI

Table 13: Rankings of the selected translations for English→Serbian according to different scores.

and Shy have around 3.5% idiomatically translated words, while all other systems have around 1% or less.

Table 13 presents rankings of the systems according to different scores: official overall score and error rate, additional overall score and error rate, as well as reward rate and total span-based score. It can be noted that Gemini, Shy, GPT and Human are almost always on the top, while CUNI is always the last. Furthermore, Gemini clearly surpasses human translation in terms of both scores and error rates in both evaluations.

Claude and ONLINE-B might be over-estimated in the official evaluation, since in the additional one they obtained notably lower scores, higher error rates, and almost no rewards, being comparable to possibly under-estimated UvA-MT. Also, human translation might be under-estimated in the official evaluation, but it is clearly worse than Gemini and comparable with Shy and GPT.

Furthermore, the middle-ranked GemTrans system turned out to be very interesting, since it generated a notable amount of idiomatic translations (5.67%), even more than humans (4.19%), but also exhibits relatively high number of errors (9.6%), many of them being morphological/agreement issues which were typical for statistical systems.

According to idiomatic translations, GemTrans and human are ranked the best, followed by the two overall top-ranked Gemini and Shy. And according to the total rate, taking into account both errors and rewards, the best three translations are Gemini, Shy and Human, followed by GemTrans and GPT.

It might be worth noting that there were 15 translations (originating from 10 source segments) without any errors: 7 were translated by Gemini, 5 by Shy, and 3 by human translators. Three of them are from the literary domain, one from speech and 11 from the social domain. Of those, 11 also have idiomatic translations: 7 by Gemini, 2 by Shy, and 2 by human translators.

As for different domains, there are no notable differences regarding idiomatic translations, although there are slightly more in news and literature (3.2% and 2.2%) than in speech and social (2.0% and 1.5%).

8 Test Suites Sub-task: "Help us break LLMs vol. 2"

For a second year in a row, we have invited the community to submit test suites in the sub-task under the call "help us break LLM". The aim is again to demonstrate evaluation methods that can expose weaknesses in LLMs which cannot be detected using standard evaluation methods. With more LLMs participating this year, and the technology advancing quickly, this call remains particularly relevant.

8.1 Setup of the Sub-task

Each test suite is a customised extension of the standard test sets that focuses on specific aspects of the machine translation (MT) output. Evaluation of the MT output takes place in a decentralised manner. Test suite providers were invited to submit their customised test sets following the setup introduced at the Third Conference on Machine Translation (Bojar et al., 2018). For this purpose, each test suite provider submitted a source-side test set, which the organisers of the General MT Shared Task then appended to their standard test sets. After generating the corresponding system outputs, the organizers returned them to the respec-

tive providers, who then conducted the evaluation according to their own methodological approach. Detailed results and analyses for each test suite are presented in separate description papers, while a consolidated summary is provided below.

8.2 Submissions

Six test suites are participating this year, covering a wide range of translation phenomena, domains, and language pairs. An overview of the test suites can be seen in Table 14. Descriptions of each submission, along with their main findings, can be found below.

DFKI (Manakhimova et al., 2025). This test suite offers a fine-grained linguistically motivated analysis of the shared task MT outputs for English–Russian, based on 465 manually devised test items, which cover 55 phenomena in 13 categories. Extending their previous test suite submissions (e.g. Avramidis et al., 2020; Macketanz et al., 2021, 2022; Manakhimova et al., 2023, 2024), the submission of this year analyzes how English–Russian machine translation (MT) systems submitted to WMT25 perform on linguistically challenging translation tasks, similar to problems used in university translator training.

The findings show that in 2025, even top-performing MT systems still struggle with translation problems that require deep understanding and rephrasing, much like human novices do. The best systems exhibit marked improvements in handling such 'extra-credit' challenges, often producing more natural translations rather than producing word-for-word renditions. However, persistent structural and lexical problems remain: literal word order carry-overs, misused verb forms, and rigid phrase translations were common, mirroring errors typically seen in beginner translator assignments.

EAA-Terminology (Hauksdottir and Steingrimsson, 2025). The EEA terminology test suite is a novel evaluation set designed to assess the capabilities of machine translation (MT) systems in handling terminology found in the EEA Agreement. It is designed for English-to-Icelandic translations, but can be easily adapted for other languages. The test suite evaluates four subdomains of the terminology in EEA regulations: science, technology, finance, and society. The test suite consists of 300 text examples in the form of sentences in English, stored in a single text file, which is to be translated by the MT systems. The suite also contains a gold standard translation meant for comparison, where each example has been translated as expected into Icelandic.

GENDER1PERSON (Popović and Lapshinova-Koltunski, 2025). The GENDER1PERSON test suite is designed for measuring gender bias in translating first-person singular forms from English into two Slavic languages, Russian and Serbian. The test suite consists of 1 000 Amazon product reviews, uniformly distributed over 10 different product categories. The bias is measured through a gender score ranging from -100 (all reviews are feminine) to 100 (all reviews are masculine).

The test suite shows that the majority of the systems participating in the WMT-2025 task for these two target languages prefer the masculine writer's gender. There is no single system which is biased towards the feminine variant. Furthermore, for each language pair, there are seven systems which are considered balanced, having the gender scores between -10 and 10.

Finally, the analysis of different products showed that the choice of the writer's gender depends to a large extent on the product. Moreover, it is demonstrated that even the systems with overall balanced scores are actually biased, but in different ways for different product categories.

IITP-legal (Singh et al., 2025a). The study critically examines various Machine Translation systems, particularly focusing on Large Language Models, using the WMT25 Legal Domain Test Suite for translating English into Hindi. It utilizes a dataset of 5, sentences designed to capture the complexity of legal texts, based on word frequency ranges from 5 to 54. Each frequency range contains 100 sentences, collectively forming a corpus that spans from simple legal terms to intricate legal provisions. Six metrics were used to evaluate the performance of the system: BLEU, METEOR, TER, CHRF++, BERTScore and COMET. The findings reveal diverse capabilities and limitations of LLM architectures in handling complex legal texts. Notably, Gemini-2.5-Pro, Claude-4 and Llama-4-Maverick topped the performance charts, showcasing the potential of LLMs for nuanced translation. Despite these advances, the study identified areas for further research, especially in improving robustness, reliability, and explainability for use in critical legal contexts. The study also supports the WMT25 subtask focused on evaluat-

Test suite	Focus	Language pair	Segments
DFKI (Manakhimova et al., 2025)	linguistic phenomena	en→ru	5,553
EEA Terminology (Hauksdottir and Steingrimsson, 2025)	legal domain	en→is	256
GENDER|PERSON (Popović and Lapshinova-Koltunski, 2025)	gender choice and agreement	en→{ru, sr}	2,000
IITP-legal (Singh et al., 2025a)	legal domain	en→hi	5,000
SportsEval (Sigurdsson et al., 2025)	sports domain	en→is	300
RoCS-MT v2 (Bawden and Sagot, 2025)	non-standard user-generated texts	en→{ar, bho, cs, et, is, ja, ko, mas, ru, sr, uk, zh}	59,340

Table 14: Overview of the participating test suites.

ing weaknesses of large language models (LLMs). The dataset and related resources are publicly available.[37]

RoCS-MT v2 (Bawden and Sagot, 2025). Robust Challenge Set for Machine Translation,[38] is designed to test MT systems' ability to translate user-generated content with non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. The original English Reddit texts are associated with manual normalisations and translations in five languages (French, German, Czech, Ukrainian and Russian). This second version of the test suite presents several improvements over the previously published version (Bawden and Sagot, 2023)., including (i) minor corrections of normalisation, (ii) corrections to reference translations and addition of alternative references to accommodate for different possible genders (e.g. of speakers) and (iii) a redesign and re-annotation of normalisation spans for further analysis of different non-standard UGC phenomena.

In previous editions of the shared task, we saw that non-standard UGC phenomena still posed problems for many models, although some of the larger, closer-sourced models handle them better. The behaviour of the systems varies greatly, with different handling of the translation of the phenomena, some systems producing more standardised outputs than others. In this edition, we saw that there was still a wide range in behaviour of systems, not all of which is accurately characterised by the automatic metric used in evaluation (COMET-Kiwi). We show some preliminary analysis showing that for elongation as a mark of expressivity (e.g. *mooooorreeee* instead of *more*), some systems are rewarded for copying the source text rather than translating the word, either in its standard or expressive form. This especially reveals issues in the current evaluation protocol for the test suite, and will require clarifying in future work.

SportsEval (Sigurdsson et al., 2025). Sports are a consistently popular domain in most news media. Although many sports attract extensive media attention and feature a rich, polysemous language, often shaped by active neologism and community-driven translations, the sports domain has received relatively little focus in MT research.

The SportsEval test suite was developed to examine the robustness of MT systems in translating sports-related texts from English into Icelandic. It covers five sports that are popular in Iceland: football (100 segments), basketball (100), chess (50), gymnastics (25), and golf (25), with a total of 300 segments. Each of these sports has a long-established domain-specific vocabulary in Icelandic, and mistranslations can easily render a text unintelligible. The segments range from single to multi-sentence passages, and most include multiple terms. While the majority are drawn from authentic usage examples, some have been adapted for brevity, and a small number are synthetic, created to illustrate specific terminological challenges.

In total, the test suite contains 971 term instances. Since some terms recur across segments, the design also enables an evaluation of translation consistency. The findings of our study indicate that current MT systems face considerable challenges in this domain.

8.3 Aggregated Results

In order to have a more general overview of the comparative system performance with regard to the test suites, we present the ranks produced by

[37] github.com/wmt25testsuite/wmt25

[38] github.com/rbawden/RoCS-MT, huggingface.co/datasets/rbawden/RoCS-MT-v2.

	WMT25 human	EEA term	sports eval	RoCS MT-v2
Gemini-2.5-Pro	1	3	1	2
Erlendur	2	2	3	4
GPT-4.1	2	10	4	3
Shy-hunyuan-MT	4	6	2	1
TowerPlus-9B	5	8	7	6
ONLINE-B	6	5	6	5
TranssionTranslate	7	4	5	9
Claude-4	7	6	9	12
TowerPlus-72B	7	12	25	8
hybrid	–	8	8	–
AMI	10	16	11	11
GemTrans	11	21	16	7
SalamandraTA	12	11	17	13
UvA-MT	12	22	26	21
CommandA	13	26	23	30
ONLINE-G	–	1	9	31
NLLB	15	14	14	24
Gemma-3-27B	–	19	12	15
IRB-MT	16	24	19	23
DeepSeek-V3	–	15	20	14
Llama-4-Maverick	–	18	13	20
Gemma-3-12B	17	22	18	25
IR-MultiagentMT	–	12	31	10
CommandA-WMT	–	16	21	16
Llama-3.1-8B	18	27	22	32
Mistral-Medium	–	20	15	22
Qwen3-235B	–	25	24	29
AyaExpanse-32B	–	28	27	33
EuroLLM-22B	–	29	30	34
CommandR-7B	–	29	31	36
Qwen2.5-7B	–	32	27	38
Mistral-7B	–	32	29	37
AyaExpanse-8B	–	29	33	39
EuroLLM-9B	–	32	34	35

Table 15: Aggregated system ranking for English→ Icelandic according to human evaluation and test suites.

	WMT25 human	dfki	gender 1person	RoCS MT-v2
Gemini-2.5-Pro	1	1	4	13
Shy-hunyuan-MT	2	5	7	1
GPT-4.1	3	5	16	8
Wenyiil	3	1	6	30
Claude-4	3	5	14	23
Algharb	5	1	1	18
DeepSeek-V3	6	5	17	10
CommandA-WMT	6	9	34	6
Yandex	8	1	3	4
UvA-MT	9	10	20	7
Qwen3-235B	10	–	22	17
Yolu	11	–	2	5
IRB-MT	11	–	12	12
GemTrans	12	–	9	2
hybrid	–	–	13	–
TowerPlus-9B	–	–	11	16
Gemma-3-27B	14	–	24	11
Gemma-3-12B	–	–	15	14
Laniqo	16	–	8	3
SRPOL	16	–	32	24
RuZh	16	–	–	–
DLUT_GTCOM	–	–	19	15
SalamandraTA	–	–	10	25
ONLINE-G	–	–	5	33
IR-MultiagentMT	–	–	30	9
AyaExpanse-32B	–	–	21	20
TowerPlus-72B	–	–	26	19
CommandA	–	–	25	21
ONLINE-W	–	–	18	31
EuroLLM-22B	–	–	23	26
TranssionTranslate	–	–	27	28
Llama-4-Maverick	–	–	33	22
AyaExpanse-8B	–	–	29	27
ONLINE-B	–	–	28	29
Qwen2.5-7B	–	–	31	34
EuroLLM-9B	–	–	37	32
Llama-3.1-8B	–	–	36	35
CommandR	–	–	35	40
NLLB	–	–	38	39
Mistral-7B	–	–	39	41
TranssionMT	–	–	40	42

Table 16: Aggregated system ranking for English→ Russian according to human evaluation and test suites.

test suites of the same language direction side by side, including the human ranks of the official WMT25 General MT test set (first column). In Tables 15, 16, 17 we present results for the language directions where we have more than one test suite. For visualisation purposes, the table rows are ordered primarily by the human ranks of the WMT25 General MT test set and then by the average of the rest of the test suites. It must be noted that this visualisation has to be taken with a grain of salt, as test suites employ different evaluation methods over different test sets off different sizes. Also, due to the different methods, the confidence intervals between the ranks have not been always taken into consideration.

One can see that there is quite some variety between the ranks of the WMT25 General MT test set and the test suites, with most obvious the ones of RoCS-MT-v2, indicating the non-standard nature of the data means that some systems which otherwise perform good, are unusually poor. The gender1person testsuite also indicates that systems with high overall performance indicate a strong male bias towards the selection of the male gender. The linguistically motivated test suite by *dfki* also has quite some variability for the systems that are evaluated. Meanwhile, the two terminology test suites are a bit closer to the WMT25 General MT testset, albeit with a few exceptions.

8.4 Summary

The evaluation of multiple test suites across diverse language pairs and domains reveals persistent challenges for current MT systems. Fine-grained linguistic analysis for English–Russian indicates high

	WMT25 human	gender 1person	RoCS MT-v2
Gemini-2.5-Pro	1	1	10
Shy-hunyuan-MT	2	7	1
GPT-4.1	2	9	6
Yolu	–	4	2
ONLINE-B	3	3	26
Claude-4	5	10	15
Human	6	–	–
TranssionTranslate	7	35	34
GemTrans	8	5	3
Algharb	–	2	14
DeepSeek-V3	8	18	9
UvA-MT	10	17	5
IRB-MT	10	13	11
DLUT_GTCOM	10	–	–
SalamandraTA	11	28	4
CommandA-WMT	11	30	25
Wenyiil	–	6	18
hybrid	–	12	–
Gemma-3-27B	–	16	8
Gemma-3-12B	13	14	12
EuroLLM-22B	–	11	17
CUNI-SFT	16	8	23
IR-MultiagentMT	–	25	7
Llama-3.1-8B	16	23	24
AyaExpanse-32B	–	15	21
NLLB	18	–	38
EuroLLM-9B	19	26	22
CommandA	–	24	16
Qwen3-235B	–	21	20
Llama-4-Maverick	–	29	13
TowerPlus-9B	–	19	29
AyaExpanse-8B	–	20	30
TowerPlus-72B	–	31	19
CommandR7B	–	22	33
Qwen2.5-7B	–	27	32
Mistral-7B	–	32	28
TranssionMT	–	34	27
ONLINE-G	–	33	31

Table 17: Aggregated system ranking for English→ Serbian according to human evaluation and test suites.

error rates in semantic roles, domain-specific terminology, and proper names, although an increase in gender-inclusive renderings was observed compared to previous years. Domain-specific evaluations for English–Icelandic, including EEA terminology and sports-related texts, demonstrate substantial difficulties in maintaining terminological accuracy and consistency. Gender bias assessment for Russian and Serbian shows a systematic preference for masculine forms, with product-specific variation even among systems classified as balanced. Legal-domain evaluation for English–Hindi confirms the superior performance of advanced LLMs such as Gemini-2.5-Pro and Claude-4, while highlighting the need for improved robustness and explainability in critical applications. Finally, robustness testing on user-generated content underscores ongoing weaknesses in handling non-standard linguistic phenomena, despite incremental progress in larger models.

9 Conclusions

The WMT 2025 General Machine Translation Task covered 30 language pairs, with human evaluation conducted on half of them across four to five domains. We prepared more challenging test set by utilizing novel difficulty sampling.

We evaluated 60 systems in total: 36 participant submissions and 24 systems collected from LLMs and popular online providers. Participation continued to grow compared to last year, and most teams utilize LLMs, often via fine-tuning.

We adopt ESA and MQM as the human evaluation protocols, which show the weak-points of models, especially in Egyptian Arabic dialect or in extremely low-resource language Maasai.

Automatic rankings did not always match human judgments: systems that topped automated metrics such as Shy-hunyuan-MT, did not consistently win under human evaluation, pointing to persistent metric bias in MBR and reinforcing that human evaluation should remain the final arbiter of translation quality.

Domain analyses showed speech as the most challenging (likely due to ASR noise), while literary was the easiest among those tested. Targeted test suites revealed remaining weaknesses in robustness to non-standard input, linguistic complexity, domain terminology, and gender choice/agreement, even as advanced LLMs improved inclusivity and performance in some specialized areas.

All source data, system outputs, and human judgments are released to support transparency, reproducibility, and further research on machine translation.

10 Limitations

We tested the general capabilities of MT systems. However, we have simplified this approach and only used three to five domains. Out of various possible modalities, we used audio and text.

Some models used pretrained metrics such as xComet or MetricX during their training, for example, using Minimum Bayes Risk or as a reward model. This significantly affected the automatic evaluation of such models giving them artificially higher scores. Furthermore, automatic metrics are limited, brittle, and biased (Karpinska et al., 2022;

Moghe et al., 2025), especially in novel domains (Zouhar et al., 2024a,b), which motivates them being superseded by human evaluation. Another potential problem may have been that test sets we use are paragraph-level; automatic metrics have usually been tested in a sentence-level scenario. Therefore, we strongly advise careful interpretation of automatic scores.

Although we use human judgments as the gold standard, giving us more reliable signal than automatic metrics, we should mention that human annotations are noisy (Wei and Jia, 2021) and their performance is affected by the quality of other evaluated systems (Mathur et al., 2020). Lastly, different annotators use different ranking strategies, which may have an effect on the system ranking.

11 Ethical Considerations

Inappropriate, controversial, and explicit content was filtered out prior to translation, keeping in mind the translators and not exposing them to such content or obliging them to translate it.

Human evaluation using Appraise for the collection of human judgements was fully anonymous. Automatically generated accounts associated with annotation tasks with single-sign-on URLs were distributed among pools of annotators and we do not store any personal information. We do store the mapping between which annotator (pseudoanonymized) annotated which segments. Annotators received standard professional translator's or evaluator's wage with respect to their countries.

Acknowledgments

This report would not have been possible without the partnership with Árni Magnússon Institute for Icelandic Studies, Charles University, Cohere, Custom.MT, Dubformer, Gates Foundation, Google, Institute of the Estonian Language, Microsoft, NTT, Toloka AI, University of Tartu, University of Tokyo. Furthermore, we are grateful to Toshiaki Nakazawa, Michael Karani and Youssef Nafea.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

Barry Haddow's participation was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant numbers 10052546 and 10039436].

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR under the project MaTOS - "ANR-22-CE23-0033-03" and as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

Martin Popel's participation was funded by TAČR grant EdUKate (TQ01000458).

Ondřej Bojar acknowledges the support by the grant CZ.02.01.01/00/23_020/0008518 (Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím"). The reference translations and manual evaluations were also supported National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO.

This work has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine translation in the era of large language models:a survey of historical and emerging problems. *Information*, 16(9).

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4555–4567. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2025. RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170. European Association for Machine Translation.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303. Association for Computational Linguistics.

Vicent Briva-Iglesias. 2025. Are AI agents the new machine translation frontier? challenges and opportunities of single- and multi-agent systems for multilingual digital communication.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the*

Fourth Workshop on Statistical Machine Translation, pages 1–28. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64. Association for Computational Linguistics.

Marine Carpuat, Omri Asscher, Kalika Bali, Luisa Bentivogli, Frédéric Blain, Lynne Bowker, Monojit Choudhury, Hal Daumé III, Kevin Duh, Ge Gao, Alvin Grissom II, Marzena Karpinska, Elaine C. Khoong, William D. Lewis, André F. T. Martins, Mary Nurminen, Douglas W. Oard, Maja Popovic, Michel Simard, and François Yvon. 2025. An interdisciplinary approach to human-centered machine translation.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268. European Association for Machine Translation.

Aaron Chatterji, Thomas Cunningham, David Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use ChatGPT.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284. Association for Computational Linguistics.

Adam Dobrowolski, Paweł Przewłocki, Paweł Przybysz, Marcin Szymański, and Dawid Siwicki. 2025. A* decoding for Machine Translation in LLMs - SRPOL participation in WMT2025. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565. ELRA and ICCL.

Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, and David Vilar. 2025. Google Translate's Research Submission to WMT2025. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.

Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt, Carlos Escolano, and Maite Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC Submission for WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765. European Language Resources Association (ELRA).

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. Salamandra technical report.

Thamme Gowda, Roman Grundkiewicz, Elijah Rippeth, Matt Post, and Marcin Junczys-Dowmunt. 2024. PyMarian: Fast neural machine translation and evaluation in python. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 328–335. Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th*

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 306–316. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Association for Computational Linguistics.

Cristian Grozea and Oleg Verbitsky. 2025. Evaluation of QWEN-3 for English to Ukrainian Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Ivan Grubišić and Damir Korencic. 2025. IRB-MT at WMT25 Translation Task: A Simple Agentic System Using an Off-the-Shelf LLM. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kamil Guttmann, Zofia Rostek, Adrian Charkiewicz, Antoni Solarski, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 General Translation Task: Self-Improved and Retrieval-Augmented Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Selma Dis Hauksdottir and Steinthor Steingrimsson. 2025. Automated Evaluation for Terminology Translation related to the EEA Agreement. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Svanhvít Lilja Ingólfsdóttir, Petur Ragnarsson, Haukur Jónsson, Haukur Simonarson, Vilhjalmur Thorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-level grammatical error correction using synthetic and curated corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316. Association for Computational Linguistics.

Svanhvít Lilja Ingólfsdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Þórðarson, and Þorvaldur Páll Helgason. 2025. Miðeind at WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Koichi Iwakawa, Keito Kudo, Subaru Kimura, Takumi Ito, and Jun Suzuki. 2025. KIKIS at WMT 2025 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Atli Jasonarson and Steinthor Steingrimsson. 2025. AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama models for Low Resource Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Josef Jon, Miroslav Hrabal, Martin Popel, and Ondřej Bojar. 2025. CUNI at WMT25 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2017a. Fasttext.zip: Compressing text classification models. In *International Conference on Learning Representations (ICLR)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017b. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.

Nikolay Karpachev, Ekaterina Enikeeva, Dmitry Popov, Arsenii Bulgakov, Daniil Panteleev, Dmitrii Ulianov, Artem Kryukov, and Artem Mekhraliev. 2025. Yandex Submission to the WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561. Association for Computational Linguistics.

Ahrii Kim. 2025a. Multi-agentMT: Deploying AI Agent in the WMT25 Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Ahrii Kim. 2025b. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165. Association for Computational Linguistics.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh

Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025a. Command-A-Translate: Raising the Bar of Machine Translation with Difficulty Filtering. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of WMT25 general machine translation systems.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Kanojia Diptesh, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Zheng Li. 2025. HYT at WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014a. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172. European Association for Machine Translation.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014b. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–Machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947. European Language Resources Association.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems

for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073. Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245. Association for Computational Linguistics.

Shushen Manakhimova, Ekaterina Lapshinova-Koltunski, Maria Kunilovskaya, and Eleftherios Avramidis. 2025. Fine-Grained Evaluation of English-Russian MT in 2025: Linguistic Challenges Mirroring Human Translator Training. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical report.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609. European Language Resources Association.

Graham Neubig. 2011. The Kyoto free translation task.

Lei Pang, Hanyi Mao, Quanjia Xiao, Chen Ruihan, Jingjun Zhang, Haixiao Liu, and Xiangyi Li. 2025. In2x at WMT25 Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.

Martin Popel, Jindřich Libovický, and Jindřich Helcl. 2022. CUNI systems for the WMT 22 Czech-Ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 352–357. Association for Computational Linguistics.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348. Association for Computational Linguistics.

Martin Popel, Lucie Polakova, Michal Novák, Jindřich Helcl, Jindřich Libovický, Pavel Straňák, Tomas Krabac, Jaroslava Hlavacova, Mariia Anisimova, and Tereza Chlanova. 2024. Charles translator: A machine translation system between Ukrainian and Czech. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3038–3045. ELRA and ICCL.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

Maja Popovic. 2021. On nature and causes of observed MT errors. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 163–175. Association for Machine Translation in the Americas.

Maja Popović and Ekaterina Lapshinova-Koltunski. 2025. GENDER1PERSON: Test Suite for estimating gender bias of first-person singular forms. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task.

In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361. Association for Computational Linguistics.

Hayate Shiroma. 2025. SH at WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Einar Sigurdsson, Magnús Már Magnússon, Atli Jasonarson, and Steinthor Steingrimsson. 2025. Up to Par? MT Systems Take a Shot at Sports Terminology. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025a. Evaluation of LLM for English to Hindi Legal Domain Machine Translation Systems. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Deepak Kumar, and Asif Ekbal. 2025b. Instruction-Tuned English to Bhojpuri Neural Machine Translation Using Contrastive Preference Optimization. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Shaomu Tan. 2025. Kaze-MT at WMT25 Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Cohere Team. 2025. Command a: An enterprise-ready large language model.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123. Association for Computational Linguistics.

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao, Bo Zeng, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516. Association for Computational Linguistics.

Christopher Lemmer Webber, Jessica Tallon, Erin Shepherd, Amy Guy, and Evan Prodromou. 2018. Activitypub, w3c recommendation. Technical report, W3C.

Johnny Wei and Robin Jia. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6840–6854. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods.

Di Wu, Yan Meng, Maya Konstantinovna Nachesa, Seth Aycock, and Christof Monz. 2025. UvA-MT's Participation in the WMT25 General Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Jiafeng Xiong and Yuting Zhao. 2025. KYUoM's Submissions to the WMT 2025 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2022. MultiWOZ 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 351–360. Association for Computational Linguistics.

Zhang Yin, Hiroyuki Deguchi, Haruto Azami, Guanyu Ouyang, Kosei Buma, Yingyi Fu, Katsuki Chousa, and Takehito Utsuro. 2025. NTTSU at WMT2025 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Dakun Zhang, Yara Khater, Ramzi Rahli, Anna Rebollo, and Josep Crego. 2025. SYSTRAN @ WMT 2025 General Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534. European Language Resources Association (ELRA).

Hao Zong, Chao Bei, Wentao Chen, Conghu Yuan, Huan Liu, and Degen Huang. 2025. DLUT and GTCOM's Large Language Model Based Translation System for WMT25. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. How to select datapoints for efficient human evaluation of NLG models?

A Error Span Annotation Miscellaneous

The following instructions were shown in the annotation interface and could be accessed at any time.
Higlighting errors: *Select the part of translation where you have identified a translation error (drag or click start & end). Click on the highlight to change error severity (minor/major) or remove the highlight.*
Choose error severity:
- *Minor errors: Style, grammar, word choice could be better or more natural.*
- *Major errors:: The meaning is changed significantly and/or the part is really hard to understand.*

Tips:
- *Missing content: If something is missing, highlight the word [MISSING] to mark the error.*
- *Tip: Highlight the word or general area of the error (it doesn't need to be exact). Use multiple highlights for different errors.*
- *Tip: Pay particular attention to translation consistency between texts across the whole document.*
- *Tip: If the translation is in the wrong language, mark it fully and assign it 0*
- *Tip: If the translation contains additional text (e.g. "Here is the translation") or alternative secondary translation, mark it as a major error.*
- *Using external tools for annotations (chatbots, LLMs) is not allowed.*

Score the translation: *After marking errors, please use the slider and set an overall score based on meaning preservation and general quality:*
- *0: Broken/poor translation.*
- *33%: Flawed: significant issues*
- *66%: Good: insignificant issues with grammar, fluency, or consistency*
- *100%: Perfect: meaning and style aligned completely with the source*

Changes to ESA interface. We introduced the following changes to the ESA interface since the previous use in WMT 2024 (Kocmi et al., 2024a):
- Dropped the requirement for a task being strictly 100 segments.
- We added a crude character-based alignment (Figure 1, top).
- We include images on the source text (Figure 1, bottom right).
- We use a translated version of the tutorial for each language pair.
- Updated the annotation instructions and the annotation scale.
- Updated the interface style slightly (Figure 1).

B Translator Brief

The following instructions were given to human translators:
In this project we wish to translate data from several domains for use in the evaluation of Machine Translation (MT). The translations produced by you will be compared against the translations produced by a variety of different MT systems. They will be released to the research community to provide a benchmark, or "gold-standard" measure for translation quality. The translation therefore needs to be a high-quality rendering of the source text into the target language, as if it was originally written directly in the target language. However, there are some constraints imposed by the intended usage:

- *All translations must be "from scratch," without post-editing from machine translation or usage of CAT tools. Post-editing machine translation would bias the evaluation, so we need to avoid it. We can detect post-editing and will reject translations that are post-edited.*
- *Translators must preserve paragraph boundaries, which are marked by empty lines in the source text, but they are free to adjust the number of sentences within each paragraph.*
- *Translators should avoid inserting parenthetical explanations into the translated text and obviously avoid losing any pieces of information from the source text. We will check the translations for quality and will reject translations that contain errors.*
- *If the original data contains errors, typos, or other problems, do not change the source sentences, instead try to prepare a correct translation as if the error wouldn't be in the source.*

- *The data contain four domains (news, speech, social, literary), each folder containing one domain source and each domain needing a specific handling*

The source files will be delivered as text files (sometimes known as "notepad" files), with paragraphs separated by an empty line. We need the translations to be returned in the same format. The translation file needs to have the same name as the original file.

Speech Domain. Your task is to translate the speech from provided video. We also provide you with automated transcription, which is not human edited and contains errors, thus should be used only as a guideline for translation from the video. Each file represents one segment of video. Videos correspond to different domains: they differ in formality, style, topics and number of speakers. The idea is to translate using the most similar language in the target language, matching as best as possible the characteristics of the source video.

Social Domain. The texts are from the social network Mastodon (similar to Twitter). Each file represents a thread or part of a thread from one or several users. Different posts within a thread are separated with empty line. Individual posts can also span several lines. The sentences have been selected so that they do not contain offensive or sensitive content (hate speech, taking-drugs, suicide, politically sensitive topics, etc.). However, profanities were kept as they were taken to be illustrative of the sociolect of online language. If however, you do not feel comfortable with translating something, please let us know.

The texts are particular in that they may contain spelling errors, slang, acronyms, marks of expressivity, etc. The idea is to translate using the most natural language in the target language, matching as best as possible the style and familiarity of the source text.

- *Spelling mistakes should not be preserved in their translations, i.e. the translation should be spelt correctly. Introduce proper capitalization in translations.*
- *Marks of expressivity (e.g. asterisks *wow*, capitals letters WOW) should be conserved as best as possible. However, do not attempt to reproduce repeated characters (e.g. woooow) in translation, as the choice as to which character to repeat is often arbitrary.*
- *There will be abbreviations and acronyms (e.g. btw -> by the way, fwiw -> for what it's worth). These do not need to be translated using abbreviation or acronyms unless an abbreviation/acronym is the best translation choice in the target language.*
- *Users (@user123) and URLs should be left as they are, i.e. not translated.*
- *Platform-specific elements such as hashtags should be translated as hashtags, but the content should be translated appropriately into the target language.*
- *Punctuation can be added if it necessary to avoid comprehension difficulties. Otherwise, we recommend following the punctuation of the source text.*

Please always refer to the screenshots included alongside the source texts. These screenshots show the original context of the entire thread contained in the source text file and should be consulted during translation. Screenshot files have the same name as the corresponding source text files but with a .png image extension instead of .txt. The screenshots may also contain images attached in the thread that can provide further context for the translation.

Literary domain. The texts in this domain are stories written by aspiring writers. Each story should be translated as one coherent text, preserving characters' speech patterns and personalities consistently. Aim to maintain the original tone and register, retaining the emotional depth of the story. Dialogues should sound natural and follow the conventions of the target language.

C System Submission Summaries

This section lists all the submissions to the translation task and provides the authors' descriptions of their submission.

C.1 Algharb (Wang et al., 2025)

In this paper, we introduce a large language model system for translation, developed through a comprehensive training pipeline. Our submissions include translations from English to Chinese, Arabic, Czech, Japanese, Korean, Russian, Ukrainian, Serbian, Bhojpuri, and Estonian, as well as from Czech to German/Ukrainian and Japanese to Chinese. Our approach integrates Machine Translation-based Supervised Fine-Tuning with post-training reinforcement via Group Relative Policy Optimization (GRPO). For decoding, we employ a Minimum Bayes Risk (MBR) algorithm enhanced with a finetuned reranker. This combined strategy ensures the generation of robust and consistent high-quality translations across a diverse set of languages.

The model is available on Hugging Face: huggingface.co/AIDC-AI/Marco-MT-Algharb.

C.2 AMI (Jasonarson and Steingrimsson, 2025)

We present the submission of the Árni Magnússon Institute's team for the WMT25 General translation task. We focus on the English→Icelandic translation direction. We pre-train Llama 3.2 3B on 10B tokens of English and Icelandic texts and fine-tune on parallel corpora. Multiple translation hypotheses are produced first by the fine-tuned model, and then more hypotheses are added by that same model further tuned using contrastive preference optimization. The hypotheses are then post-processed using a grammar correction model and post-processing rules before the final translation is selected using minimum Bayes risk decoding. We found that while it is possible to generate translations of decent quality based on a lightweight model with simple approaches such as the ones we apply, our models are quite far behind the best participating systems and it would probably take somewhat larger models to reach competitive levels.

The model is available on Hugging Face: huggingface.co/arnastofnun/Llama-3.2-3B-wmt25-AMI-en-is.

C.3 CGFOKUS (Grozea and Verbitsky, 2025)

We report here the outcome of evaluating Qwen3 for the English to Ukrainian language pair of the general MT task of WMT 2025. In addition to the quantitative evaluation, a qualitative evaluation was performed, leveraging the cooperation with a native Ukrainian speaker - therefore we present an example-heavy analysis of the typical failures the LLMs still do when translating natural language, particularly into Ukrainian. We report also on the practicalities of using LLMs, such as on the difficulties of making them follow instruction, on ways to exploit the increased "smartness" of the reasoning models while simultaneously avoiding the reasoning part interfering wrongly with the chain of which the LLM is just one element.

C.4 COILD-BHO (Singh et al., 2025b)

This paper presents an English to Bhojpuri machine translation (MT) system developed for the WMT25 General MT Shared Task. Given the low-resource nature of Bhojpuri, we adopt a two-stage training pipeline: unsupervised pretraining followed by supervised fine-tuning. During pretraining, we use a 300,000-sentence corpus comprising 70% Bhojpuri monolingual data and 30% English data to establish language grounding. The fine-tuning stage utilizes 29,749 bilingual English to Bhojpuri sentence pairs (including training, validation, and test sets). To adapt the system to instruction-following scenarios, we apply a novel optimization strategy: Contrastive Preference Optimization (CPO). This technique enables the model to capture fine-grained translation preferences and maintain semantic fidelity in instruction-tuned settings. We evaluate our system across multiple metrics, demonstrating moderate performance in low-resource MT tasks, particularly in diverse domains such as literary, news, social, and speech.

The model is available at: drive.google.com/drive/folders/1ZzJ9ZlfqaT-fEo5umovNN4HWkZhOlqqC.

C.5 CommandA-WMT (Kocmi et al., 2025a)

We built our system on top of Command-A using a direct preference optimization with data preparation pipeline that emphasizes robust data quality control, primarily incorporating standard quality filtering along with a novel difficulty filtering component, which serves as the key innovation of our approach. The final translation is built through step-by-step reasoning, and we employ limited Minimum Bayes Risk decoding with a limited candidate pool size of 20, using MetricX-XL as the primary utility metric. For unsupported languages, we use a second model prepared identically but with an additional initial supervised fine-tuning step for the unsupported languages that Command-A model has not been trained on.

The model is available on Hugging Face: huggingface.co/CohereLabs/command-a-translate-08-2025

C.6 CUNI-EdUKate-v1 (Jon et al., 2025)

CUNI-EdUKate-v1 is an unconstrained system trained on educational domain data using LoRA, SFT, and Contrastive Preference Optimization. It is also fine-tuned from the EuroLLM-9B-Instruct model. It only supports the cs2uk language direction and, unlike CUNI-MH-v2, both training and inference were done at the sentence level.

C.7 CUNI-MH-v2 (Jon et al., 2025)

CUNI-MH-v2 is a constrained system trained on partially synthetic data sampled from the CzEng 2.0 dataset using LoRA and Contrastive Preference Optimization. We plan on releasing both the model and the filtered training data. It is fine-tuned from the EuroLLM-9B-Instruct model. We currently only support two language directions, en2cs and cs2de, and offer separate LoRA adapters for each. The translations were done on the paragraph level.

The models are available on Hugging Face: huggingface.co/hrabalm/CUNI-MH-v2-encs and huggingface.co/hrabalm/CUNI-MH-v2-csde.

C.8 CUNI-SFT (Jon et al., 2025)

This paper describes the joint effort of Phrase a.s. and CUNI/UFAL on the WMT25 Automated Translation Quality Evaluation Systems Shared Task. Both teams participated both in a collaborative and competitive manner, i.e. they each submitted a system of their own as well as a contrastive joint system ensemble. In Task 1, we show that such an ensembling—if chosen in a clever way—can lead to a performance boost. We present the analysis of various kinds of systems comprising both "traditional" NN-based approach, as well as different flavours of LLMs—off-the-shelf commercial models, their fine-tuned versions, but also in-house, custom-trained alternative models. In Tasks 2 and 3 we show Phrase's approach to tackling the tasks via various GPT models: Error Span Annotation via the complete MQM solution using non-reasoning models (including fine-tuned versions) in Task 2, and using reasoning models in Task 3.

The model is available on Hugging Face: huggingface.co/ufal/wmt25-cuni-sft.

C.9 CUNI-Transformer and CUNI-DocTransformer (Popel et al., 2022, 2019)

CUNI-Transformer and CUNI-DocTransformer rely on standard NMT training with Block backtranslation and optionally document-level training.

The models are available at lindat.mff.cuni.cz/repository/items/b1cfdecf-fda3-4198-a537-e58a20ddea60 and lindat.mff.cuni.cz/repository/items/4b5d758f-ca9e-4ca6-8129-2331928ba950.

C.10 DLUT_GTCOM (Zong et al., 2025)

This paper presents the submission from Dalian University of Technology (DLUT) and Global Tone Communication Technology Co., Ltd. (GTCOM) to the WMT25 General Machine Translation Task. Amidst the paradigm shift from specialized encoder-decoder models to general-purpose Large Language Models (LLMs), this work conducts a systematic comparison of both approaches across five language pairs. For traditional Neural Machine Translation (NMT), we build strong baselines using deep Transformer architectures enhanced with data augmentation. For the LLM paradigm, we explore zero-shot performance

and two distinct supervised fine-tuning (SFT) strategies: direct translation and translation refinement. Our key findings reveal a significant discrepancy between lexical and semantic evaluation metrics: while strong NMT systems remain competitive in BLEU scores, fine-tuned LLMs demonstrate marked superiority in semantic fidelity as measured by COMET. Furthermore, we find that fine-tuning LLMs for direct translation is more effective than for refinement, suggesting that teaching the core task directly is preferable to correcting baseline outputs.

C.11 Erlendur (Ingólfsdóttir et al., 2025)

We present Miðeind's system contribution for English-to-Icelandic translation. We participate in the Terminology Shared Task with the same system. Erlendur is a multilingual LLM-based translation system which employs a multi-stage pipeline approach, with enhancements especially for translations from English to Icelandic. We address translation quality and grammatical accuracy challenges in current LLMs through a hybrid prompt-based approach that can benefit lower-resource language pairs. In a preparatory step, the LLM analyzes the source text and extracts key terms for lookup in an English-Icelandic dictionary. Main results of the analysis and the retrieved dictionary results are then incorporated into the translation prompt. When provided with a custom glossary, the system identifies relevant terms from the glossary and incorporates them into the translation as well, to ensure consistency in terminology. For longer inputs, the system maintains translation consistency by providing contextual information from preceding text chunks. Lastly, Icelandic target texts are passed through our custom-developed seq2seq language correction model (Ingólfsdóttir et al., 2023), where grammatical errors are corrected. Using this hybrid method, Erlendur delivers high-quality translations, without fine-tuning.

C.12 HYT (Li, 2025)

This paper illustrates the submission system of the HYT team for the WMT25 General Machine Translation shared task. We submitted translations for all translation directions in the general machine translation task and test suites subtask. The ID of our submission in OCELoT system is 43, which can be categorized as being in the unconstrained track. The base model we use is Hunyuan-TurboS. Overall, we first performed continued pretraining(CPT) using open-source data to enhance the model's multilingual capabilities. Then, we used DeepSeek-V3-03241 to synthesize a large amount of parallel data and performed Reinforcement learning on the CPT model. Finally, we used ensemble learning to further improve translation quality.

C.13 GemTrans (Finkelstein et al., 2025)

Large Language Models have shown impressive multilingual capabilities, where translation is one among many tasks. Google Translate's submission to the 2025 WMT evaluation tries to research how these models behave when pushing their translation performance to the limit. Starting with the strong Gemma 3 model, we carry out supervised fine-tuning on high quality, synthetically generated parallel data. Afterwards we perform an additional Reinforcement Learning step, with reward models based on translation metrics to push the translation capabilities even further. Controlling the combination of reward models, including reference-based and quality estimation metrics, we found that the behaviour of the model could be tailored towards a more literal or more creative translation style. Our two submissions correspond to those two models. We chose the more creative system as our primary submission, targeting a human preference for better sounding, more naturally flowing text, although at the risk of losing on the accuracy of the translation. It is an open question to find the sweet spot between these two dimensions, which certainly will depend on the specific domain to handle and user preferences.

C.14 In2x (Pang et al., 2025)

This paper presents the open-system submission by the In2x research team for the WMT25 General Machine Translation Shared Task. Our submission focuses on Japanese-related translation tasks, aiming to explore a generalizable paradigm for extending large language models (LLMs) to other languages. This paradigm encompasses aspects such as data construction methods and reward model design. The ultimate goal is to enable large language model systems to achieve exceptional performance in low-resource or less commonly spoken languages.

C.15 IR-MultiagentMT (Kim, 2025a)

We introduce our model, referred to as Multi-agentMT, for participation in the WMT 25 General Machine Translation Shared Task. This model operationalizes the notion of an AI Agent by employing a multi-agent workflow known as Prompt Chaining (Briva-Iglesias, 2025) alongside the automatic MQM (Multidimensional Quality Metrics) error annotation framework designated as RUBRIC-MQM (Kim, 2025b). Our primary submission is developed through the Translate-Postedit-Proofread paradigm, whereby the positions of the errors are clearly marked and enhanced throughout the process. Our study suggests that a semi-autonomous agent scheme in Machine Translation is viable with an older and smaller model in some language pairs, resulting in comparable results with 2.3x faster speed and only 2% of the budget.

C.16 IRB-MT (Grubišić and Korencic, 2025)

Large Language Models (LLMs) have been demonstrated to achieve state-of-art results on machine translation. LLM-based translation systems usually rely on model adaptation and fine-tuning, requiring datasets and compute. The goal of our team's participation in the "General Machine Translation" and "Multilingual" tasks of WMT25 was to evaluate the translation effectiveness of a resource-efficient solution consisting of a smaller off-the-shelf LLM coupled with a self-refine agentic workflow. Our approach requires a high-quality multilingual LLM capable of instruction following. We select Gemma3-12B among several candidates using the pretrained translation metric MetricX-24-XL and a small development dataset. WMT25 automatic evaluations place our solution in the mid tier of all WMT25 systems, and also demonstrate that it can perform competitively for approximately 16% of language pairs.

C.17 Kaze-MT (Tan, 2025)

This paper describes the Kaze-MT submission to the WMT25 General Machine Translation task for the Japanese-Chinese track. The system relies on a minimalist Test-Time Scaling (TTS) pipeline composed of three stages: Sampling, Scoring, and Selection. In the sampling stage, we utilize zero-shot Qwen 2.5 models (72B and 14B) to generate 512 candidate translations under a fixed temperature schedule, encouraging diversity without compromising fluency. In the scoring stage, each candidate is evaluated using multiple quality estimation (QE) models, namely KIWI22, MetricX-24, and ReMedy-24. Finally, we select the final candidate based on rank aggregation across QE scores. Our approach requires no fine-tuning, in-context examples, or specialized decoding heuristics, and we participate in both constrained and unconstrained tracks. Preliminary results show competitive performance on automatic metrics, with final human evaluation results to be reported in the camera-ready version.

C.18 KIKIS (Iwakawa et al., 2025)

We participated in the constrained English–Japanese track of the WMT 2025 General Machine Translation Task. Our system collected the outputs produced by multiple subsystems, each of which consisted of LLM-based translation and reranking models configured differently (e.g., prompting strategies and context sizes), and reranked those outputs. Each subsystem generated multiple segment-level candidates and iteratively selected the most probable one to construct the document translation. We then reranked the document-level outputs from all subsystems to obtain the final translation. For reranking, we adopted a text-based LLM reranking approach with a reasoning model to take long contexts into account. Additionally, we built a bilingual dictionary on the fly from the parallel corpus to make the system more robust to rare words.

C.19 KYUoM (Xiong and Zhao, 2025)

This paper describes the KYUoM team's submission system for the WMT 2025 general translation task. We focused on exploring the capabilities of inductive generalization from a multimodal domain to a text-based domain of machine translation. Our submission system consists of a two-stage adaptation process with multimodal domain learning in the first stage and textual domain adaptation in the second stage for the English to Ukrainian task in the unconstrained track. The main advance is using a GAT adapter to achieve two-stage continuous learning for cross-modal generalization.

C.20 Laniqo (Guttmann et al., 2025)

This work describes Laniqo's submission to the constrained track of the WMT25 General MT Task. We participated in 11 translation directions. Our approach combines several techniques: fine-tuning the EuroLLM-9B-Instruct model using Contrastive Preference Optimization on a synthetic dataset, applying Retrieval-Augmented Translation with human-translated data, implementing Quality-Aware Decoding, and performing postprocessing of translations with a rule-based algorithm. We analyze the contribution of each method and report improvements at every stage of our pipeline.

The model is available on Hugging Face: huggingface.co/laniqo/WMT25-EuroLLM-9B-CPO.

C.21 NTTSU (Yin et al., 2025)

This paper presents the submission of NTTSU for the constrained track of the English–Japanese and Japanese–Chinese language directions at the WMT2025 general translation task. For each translation direction, we build translation models from a large language model by combining continual pretraining, supervised fine-tuning, and preference optimization based on the translation quality and adequacy. We finally generate translations via context-aware MBR decoding to maximize translation quality and document-level consistency.

The models are available on Hugging Face: huggingface.co/UtsuroLab/WMT25_En-Ja and huggingface.co/UtsuroLab/WMT25_Ja-Zh.

C.22 RuZH-Eole (no paper submission)

Eole NLP Submission uses Tower+ 9B model with an extra layer for quality estimation. It generates multiple hypotheses and rank them according to an internal score.

C.23 SalamandraTA (Gilabert et al., 2025)

In this paper, we present the SALAMANDRA**TA** family of models, an improved iteration of SALAMANDRA LLMs (Gonzalez-Agirre et al., 2025) specifically trained to achieve strong performance in translation-related tasks for 38 European languages. SALAMANDRA**TA** comes in two scales: 2B and 7B parameters. For both versions, we applied the same training recipe with a first step of continual pre-training on parallel data, and a second step of supervised fine-tuning on high-quality instructions. The BSC submission to the WMT25 General Machine Translation shared task is based on the 7B variant of SALAMANDRA**TA**. We first adapted the model vocabulary to support the additional non-European languages included in the task. This was followed by a second phase of continual pre-training and supervised fine-tuning, carefully designed to optimize performance across all translation directions for this year's shared task. For decoding, we employed two quality-aware strategies: Minimum Bayes Risk Decoding and Tuned Re-ranking using COMET and COMET-KIWI respectively.

We publicly release both the 2B and 7B versions of SALAMANDRA**TA**, along with the newer SALAMANDRA**TA**-V2 model, on Hugging Face: huggingface.co/LangTech-MT/salamandraTA-7b-instruct-WMT25.

C.24 SH (Shiroma, 2025)

We participated in the unconstrained track of the English-to-Japanese translation task at the WMT 2025 General Machine Translation Task. Our submission leverages several large language models, all of which are trained with supervised fine-tuning, and some further optimized via preference learning. To enhance translation quality, we introduce an automatic post-editing model and perform automatic post-editing. In addition, we select the best translation from multiple candidates using Minimum Bayes Risk (MBR) decoding with the use of COMET-22 and LaBSE-based cosine similarity as evaluation metrics.

C.25 Shy-hunyuan-MT (Zheng et al., 2025)

This paper presents our submission to the WMT25 shared task on machine translation, for which we propose **S**ynergy-en**h**anced polic**y** optimization, named Shy, a novel two-phase training framework that synergistically combines ensemble knowledge distillation with reinforcement learning optimization. In the first phase, we introduce a multi-stage training framework that harnesses the complementary strengths of multiple state-of-the-art large language models to generate diverse, high-quality translation

candidates. These candidates serve as pseudo-references to guide the supervised fine-tuning of our model, Hunyuan-7B, effectively distilling the collective knowledge of multiple expert systems into a single efficient model. In the second phase, we further refine the distilled model through Group Relative Policy Optimization, a reinforcement learning technique that employs a composite reward function. By calculating reward from multiple perspectives, our model ensures better alignment with human preferences and evaluation metrics. Extensive experiments across multiple language pairs demonstrate that our model **Shy-hunyuan-MT** yields substantial improvements in translation quality compared to baseline approaches. Notably, our framework achieves competitive performance with state-of-the-art systems while maintaining computational efficiency through knowledge distillation and strategic ensemble.

The model is available on Hugging Face: huggingface.co/collections/tencent/hunyuan-mt-68b42f76d473f82798882597.

C.26 SRPOL (Dobrowolski et al., 2025)

This work presents an innovative decoding approach utilizing the A* (A-star) algorithm, which generates a diverse and precise set of translation hypotheses. Subsequent reranking through the Noisy Channel Model Reranking and Quality Estimation selects the best among these diverse hypotheses, leading to a significant improvement in translation quality. This approach achieves up to a 0.5-point reduction in the MetricX-24 score and a 1.5-point increase in the COMET score. The A* algorithm can be applied to decoding in any LLMs or classic transformers. The experiment shows that by using freely available, open-source MT models, it is possible to achieve translation quality comparable to the best online translators and LLMs using only a PC under your desk.

C.27 Systran (Zhang et al., 2025)

We present an English-to-Japanese translation system built upon the EuroLLM-9B (Martins et al., 2025) model. The training process involves two main stages: continue pretraining (CPT) and supervised fine-tuning (SFT). After both stages, we further tuned the model using a development set to optimize performance. For training data, we employed both basic filtering techniques and high-quality filtering strategies to ensure data cleanness. Additionally, we classify both the training data and development data into four different domains and we train and fine-tune with domain specific prompts during system training. Finally, we applied Minimum Bayes Risk (MBR) decoding and paragraph-level reranking for post-processing to enhance translation quality.

The models are available on Hugging Face: huggingface.co/collections/Systran/wmt25-en-ja-6867eed78ea21e28a282aaed.

C.28 TranssionMT (no paper submission)

The team employs the Transformer architecture and finetuning the translation of a specific language within the multilingual pretrained model to enhance its translation performance. They adopts various strategies, such as finetuning language model instructions, joint training of similar languages, integrated model decision-making, and non-English data mining, all aimed at improving the translation outcomes.

C.29 TranssionTranslate (no paper submission)

This paper presents our machine translation system developed for the WMT25 shared task. Our approach leverages state-of-the-art neural architectures, including transformer-based models with advanced pre-training and fine-tuning techniques. We focus on multilingual and domain-adaptive strategies to enhance translation quality across diverse language pairs. Key features include: (1) large-scale pretraining on parallel and monolingual corpora, (2) dynamic data filtering and domain adaptation, (3) ensemble and reranking methods to improve fluency and accuracy. We explore both supervised and zero-shot settings, particularly for low-resource languages. Our system demonstrates competitive performance on WMT25 evaluation benchmarks, achieving improvements in BLEU, TER, and human evaluation metrics. We analyze challenges such as rare word translation, syntactic divergence, and robustness to noisy inputs. The results highlight the effectiveness of our approach in balancing generalization and language-specific optimization. This work contributes insights into scalable and adaptive MT systems, with potential

applications in multilingual NLP tasks. Future directions include better handling of linguistic diversity and real-time adaptation.

C.30 UvA-MT (Wu et al., 2025)

The UvA-MT's submission is competing in the unconstrained track across all 16 translation directions. Unusually, this year we use only the source side of the test set to generate synthetic data for LLM training, and translations are produced using pure beam search for submission. Overall, our approach can be seen as a special variant of data distillation, motivated by two key considerations: (1) perfect domain alignment, where the training and test domains are distributionally identical; and (2) the strong teacher model, GPT-4o-mini, offers high-quality outputs as both a reliable reference and a fallback in case of mere memorization. Interestingly, the outputs of the resulting model, trained on Gemma3-12B using Best-of-N (BoN) outputs from GPT-4o-mini, outperform the original BoN outputs in some high-resource languages across various metrics, including CometKiwi-XXL which is the very metric used for BoN selection. We attribute this to a successful model ensemble, where the student model (Gemma3-12B) retains the strengths of the teacher (GPT-4o-mini) while implicitly avoiding their flaws. Our experiments on other datasets, such as WMT24++, also confirm this observation.

C.31 Wenyiil (no paper submission)

This paper introduces Wenyiil, an advanced translation system based on a large language model (LLM). This multilingual model supports 13 language directions, and its superior performance is derived from a comprehensive training process that includes multi-stage supervised fine-tuning (SFT) for translation tasks and a two-stage post-training scheme. Furthermore, we propose a novel hybrid decoding strategy to overcome the limitations of standard decoding. This method integrates word alignment with an advanced Minimum Bayes Risk (MBR) re-ranking algorithm. This approach not only enhances translation stability but also ensures excellent accuracy across diverse linguistic contexts.

C.32 Yandex (Karpachev et al., 2025)

This paper describes Yandex's submission to the WMT25 General Machine Translation task. We participate in the English-to-Russian translation direction and propose a purely LLM-based translation model. Our training procedure comprises a training pipeline of several stages built upon YandexGPT, an in-house general-purpose LLM. In particular, firstly, we employ continual pretraining (post-pretrain) for MT task for initial adaptation to multilinguality and translation. Subsequently, we use SFT on parallel document-level corpus in the form of P-Tuning. Following SFT, we propose a novel alignment scheme of two stages, the first one being a curriculum learning with difficulty schedule and a second one - training the model for tag preservation and error correction with human post-edits as training samples. Our model achieves results comparable to human reference translations on multiple domains.

C.33 Yolu (no paper submission)

This paper details Yolu's submission for the WMT'25 General Machine Translation Task. Our work, situated within the constrained track, investigates the efficacy of Reinforcement Learning (RL) in enhancing machine translation. Our system is built upon the open-source Qwen3 model. We introduce a robust methodology for continuous performance improvement, which combines meticulous data cleaning with advanced data distillation techniques. This is complemented by a multi-stage optimization strategy, sequentially employing Continued Pre-Training (CPT), Supervised Fine-Tuning (SFT), Contrastive Preference Optimization (CPO), and a novel policy optimization algorithm, Decoupled Clip and Dynamic sAmpling Policy Optimization (DAPO). Furthermore, we integrate a Quality Estimation (QE) model to facilitate online QE distillation, thereby refining the model's output during the decoding phase.

D Official Ranking Results (extends Section 7.4)

Results tables legend

The human score is the micro-average of human judgments across all domains and double annotations (single annotations for MQM language pairs). AutoRank is calculated from automatic metrics as per (Kocmi et al., 2025b). Significance testing is done using a Wilcoxon signed rank test with a p-value threshold of 5%. The rank range for the ith model begins as $\langle i, i \rangle$ and is expanded in both directions until a significant difference is found. Clusters are formed such that their constituent rank ranges do not overlap.

Systems are either constrained (white), or unconstrained (gray). Systems that do not officially support the language pair are marked with 🚫 and those where language support cannot be verified are marked with ❓. The [M] suffix marks systems (submitted by the WMT organizers) that were trained/tuned with specific MT instructions, but prompted without these specific instructions (using a generic setup, same for all LLMs, see Section 4.2), which could disadvantage these systems.

English→Arabic (Egyptian)

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	**Human**	78.5		83.8	74.7	79.1	77.3
2-2	GPT-4.1	77.0	6.7	80.4	74.7	77.3	76.2
3-3	CommandA	74.0	8.6	81.3	66.8	75.6	73.8
4-4	Gemini-2.5-Pro	60.6	5.8	88.4	0.9	84.3	80.8
5-6	DeepSeek-V3❓	56.8	7.0	66.0	33.2	64.5	69.9
5-6	Claude-4	55.7	7.8	73.5	23.7	64.5	69.5
7-7	IRB-MT	51.9	11.1	62.2	20.0	68.2	61.6
8-9	Mistral-Medium	36.0	7.7	44.2	0.1	46.4	64.9
8-9	CommandA-WMT	34.6	4.1	37.0	30.4	18.1	66.8
10-10	UvA-MT	29.0	4.2	12.4	8.2	42.0	58.8
11-14	CommandR7B	3.7	11.6	0.0	0.6	3.2	13.9
11-14	GemTrans	3.7	3.5	0.0	0.1	1.6	17.6
11-16	Algharb	3.2	2.7	0.0	1.2	1.6	12.9
11-16	Shy-hunyuan-MT	3.2	1.0	0.0	2.6	1.7	10.5
13-16	AyaExpanse-8B	2.0	9.9	0.0	0.0	1.7	8.2
12-16	ONLINE-B	1.7	6.5	0.0	0.6	1.8	5.0
17-19	Yolu	1.4	5.5	0.0	0.0	1.6	4.8
15-18	Wenyiil	1.4	2.5	0.0	0.6	1.7	3.5
19-19	SRPOL🚫	0.9	8.1	0.0	0.0	1.6	2.4
20-39	19 systems not human-evaluated		...				

English→Estonian

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	**Human**	83.1		96.8	83.0	82.3	68.2
2-2	Gemini-2.5-Pro	78.8	2.5	72.3	78.1	88.9	71.0
3-4	Wenyiil	72.6	2.6	63.5	77.2	78.3	67.8
3-4	GPT-4.1	72.2	3.0	79.0	71.4	72.2	64.9
5-6	Algharb	70.4	3.9	51.9	77.0	79.7	68.0
5-6	Shy-hunyuan-MT	70.3	1.0	71.3	73.8	69.3	65.2
7-8	ONLINE-B	60.2	6.0	80.3	52.6	58.8	49.7
7-8	Yolu	59.5	3.8	66.9	58.5	60.4	50.5
9-9	TranssionTranslate❓	57.1	7.3	55.5	59.4	64.9	42.9
10-11	Claude-4❓	53.0	6.5	51.0	53.5	58.7	45.2
10-12	GemTrans	51.7	5.1	38.6	51.0	58.3	57.6
11-14	CommandA-WMT🚫	50.1	6.1	53.7	48.7	52.0	45.2
12-15	SRPOL	49.4	5.7	40.4	53.8	54.1	46.2
12-17	Laniqo	48.6	5.2	50.1	53.7	45.8	43.5
13-17	EuroLLM-22B-pre.[M]	47.2	8.1	49.8	42.2	51.9	44.1
14-18	SalamandraTA	46.7	6.3	40.2	49.5	48.8	46.9
14-18	UvA-MT	46.4	5.9	55.0	38.5	45.4	49.7
16-18	Gemma-3-27B	45.9	7.6	32.6	51.7	46.9	51.4
19-19	IRB-MT	32.4	11.4	14.8	35.8	36.4	42.1
20-40	20 systems not human-evaluated		...				

English→Bhojpuri

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro	94.9	1.0	98.8	95.9	95.1	88.3
2-3	**Human**	92.6		97.7	92.7	94.1	83.8
2-3	Algharb	91.1	2.8	89.1	95.8	92.1	84.5
4-4	Wenyiil	90.9	2.5	94.0	93.6	91.0	82.7
5-6	Claude-4?	83.2	4.5	90.7	88.7	73.7	81.2
5-6	GPT-4.1?	82.8	5.5	76.2	92.5	84.8	73.1
7-8	TranssionTranslate?	79.5	4.3	88.6	82.1	72.0	76.9
7-10	DeepSeek-V3?	77.3	5.1	85.3	82.3	71.7	69.0
8-10	Llama-4-Maverick	76.4	6.5	71.9	78.1	76.8	78.6
8-10	NLLB	75.6	6.6	77.1	78.3	74.8	71.0
11-12	CommandA ⃠	72.6	6.5	84.6	73.0	67.8	65.2
11-12	Yolu	72.4	5.7	79.1	70.6	69.8	71.4
13-14	TranssionMT	70.1	6.2	53.7	76.0	74.4	74.2
13-15	COILD-BHO	68.7	8.9	75.9	83.3	71.6	32.7
14-15	ONLINE-B	67.2	4.1	62.0	70.5	62.6	76.0
16-16	IRB-MT	59.6	11.4	62.7	74.2	52.2	45.8
17-17	Gemma-3-27B?	56.0	8.3	42.5	47.2	65.7	69.9
18-18	SalamandraTA	35.7	12.1	27.6	35.1	44.8	31.7
19-19	Shy-hunyuan-MT	1.7	11.5	0.0	3.0	1.7	1.9
20-37	17 systems not human-evaluated		...				

English→Masai

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro?	9.8	6.1	17.5	7.8	6.3	8.7
2-2	**Human**	9.6		16.5	4.6	10.9	7.9
3-3	Claude-4?	7.7	2.6	17.0	3.5	4.4	8.2
4-6	AyaExpanse-8B ⃠	6.0	8.2	13.8	2.8	5.4	2.6
4-5	Llama-4-Maverick ⃠	5.2	3.2	2.5	8.8	3.8	4.4
6-6	Shy-hunyuan-MT ⃠	4.8	1.0	12.7	2.9	1.7	2.7
7-13	AyaExpanse-32B ⃠	3.1	7.1	0.0	2.5	5.7	4.1
4-8	DeepSeek-V3?	3.0	6.2	1.7	4.0	1.8	4.8
9-13	Llama-3.1-8B ⃠	3.0	8.1	0.1	1.4	8.1	2.1
9-13	Gemma-3-12B?	3.0	8.8	0.0	1.7	6.6	3.9
9-13	Qwen2.5-7B?	2.8	8.6	0.1	2.6	5.1	3.1
9-13	Qwen3-235B ⃠	2.7	3.0	0.2	0.8	5.7	5.3
9-13	TranssionMT	2.5	5.9	0.9	1.6	3.2	5.4
14-18	CommandR7B ⃠	1.6	4.3	0.1	0.0	3.4	3.9
14-18	CommandA-WMT ⃠	1.5	6.4	0.0	3.9	0.0	1.5
14-16	CommandA ⃠	1.3	7.9	0.1	0.1	3.0	2.7
17-18	TowerPlus-9B[M] ⃠	0.8	5.3	0.0	1.2	0.1	2.1
17-18	EuroLLM-9B[M] ⃠	0.7	8.2	0.0	0.3	1.6	1.2
19-19	EuroLLM-22B-pre.[M] ⃠	0.5	8.2	0.0	1.1	0.0	0.6
20-29	9 systems not human-evaluated		...				

English→Russian

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro	83.4	4.4	91.4	82.5	83.7	74.8
2-2	Shy-hunyuan-MT	80.2	1.0	89.8	83.4	78.5	67.5
3-5	Wenyiil	78.2	4.8	90.7	82.4	70.3	72.2
3-5	GPT-4.1	76.2	5.4	85.9	73.3	75.3	70.2
3-5	Claude-4	75.9	8.7	94.8	73.5	71.2	65.6
6-9	DeepSeek-V3?	73.6	5.7	89.0	66.5	74.8	63.0
5-8	Algharb	73.3	5.2	62.8	83.6	70.2	77.0
6-9	CommandA-WMT	73.2	4.2	92.3	72.9	71.3	54.8
8-10	Yandex	72.0	4.5	90.0	77.9	63.8	58.0
9-11	**Human**	70.5		91.5	65.9	71.8	49.5
10-12	UvA-MT	69.1	4.5	79.0	75.9	63.3	58.6
11-14	Qwen3-235B	67.6	8.8	74.5	67.0	68.6	58.5
12-15	IRB-MT	65.4	10.1	77.2	65.5	63.7	54.3
12-15	Yolu	64.5	6.9	80.9	63.4	63.6	48.0
13-16	GemTrans	62.5	5.1	51.5	79.9	59.4	56.5
15-16	Gemma-3-27B	61.7	8.9	74.5	56.5	60.4	56.3
17-19	RuZh?	57.9	9.6	54.4	58.3	65.0	48.1
17-19	SRPOL	56.9	10.6	75.4	59.8	53.0	38.2
17-19	Laniqo	56.2	8.8	58.3	56.6	57.8	50.1
20-42	22 systems not human-evaluated		...				

English→Ukrainian

Rank	System	Human	AutoRank	literary	news	social	speech
1-3	Gemini-2.5-Pro	90.3	3.3	93.8	90.5	90.2	86.2
1-3	Algharb	90.0	4.2	91.5	91.2	89.6	87.2
1-3	Wenyiil	89.5	3.5	91.9	90.9	89.9	84.1
4-5	Shy-hunyuan-MT	88.4	1.0	90.8	90.2	89.0	82.0
4-5	GemTrans	88.2	4.6	89.9	90.8	88.2	82.4
6-7	GPT-4.1	87.9	3.5	90.3	88.9	88.5	82.7
5-8	**Human**	87.3		95.2	85.3	86.3	82.7
7-9	UvA-MT	86.4	4.4	86.0	88.0	87.9	81.5
8-13	CommandA-WMT	86.3	3.9	87.1	87.1	86.4	84.0
9-13	Llama-4-Maverick	86.2	8.8	91.2	86.0	87.2	78.8
9-13	DeepSeek-V3❓	85.8	5.0	87.4	88.0	85.0	82.2
9-14	Claude-4❓	85.6	7.0	87.3	85.3	86.5	81.9
9-13	Yolu	85.4	6.0	88.0	88.3	87.7	73.8
14-16	Mistral-Medium❓	84.5	6.0	85.3	86.0	84.1	82.5
14-16	TowerPlus-9B[M]	84.2	8.8	86.3	86.4	84.7	77.6
14-16	CommandA	84.0	7.4	84.4	87.4	83.3	79.7
17-17	IRB-MT	82.9	8.2	83.8	87.1	83.4	74.8
18-19	SRPOL	79.9	8.4	76.5	83.1	84.3	71.1
18-19	Laniqo	79.8	7.7	81.2	82.2	82.0	70.6
20-44	24 systems not human-evaluated		...				

English→Italian

Rank	System	Human	AutoRank	literary	news	social	speech
1-4	Gemini-2.5-Pro	79.4	4.4	74.4	86.1	80.6	71.9
1-4	GemTrans	79.4	5.2	85.8	79.0	81.7	68.0
1-4	GPT-4.1	79.0	4.5	87.0	73.9	83.3	69.3
1-4	Shy-hunyuan-MT	78.7	1.0	74.4	80.4	83.4	71.8
5-7	CommandA-WMT	75.5	2.6	77.9	79.4	77.0	63.3
5-8	Mistral-Medium❓	73.8	7.1	79.1	67.8	79.9	65.4
5-10	CommandA	73.2	8.4	82.3	80.2	67.4	62.6
6-10	Claude-4	72.1	8.4	73.9	75.5	70.6	67.7
7-10	UvA-MT	71.8	5.3	68.4	74.1	77.5	60.7
7-10	DeepSeek-V3❓	71.7	6.1	63.6	75.7	73.8	69.9
11-11	Qwen3-235B	67.0	7.2	60.8	71.3	71.8	57.4
12-13	TowerPlus-9B[M]	61.2	11.3	71.6	62.6	57.8	53.5
12-13	IRB-MT	60.3	10.2	53.7	67.1	62.1	53.2
14-16	SalamandraTA	57.5	10.3	45.5	69.9	62.2	41.6
14-16	AyaExpanse-8B	57.0	14.9	50.8	65.8	60.5	42.9
14-16	EuroLLM-9B[M]	56.6	15.2	58.4	57.1	57.1	52.7
17-18	Gemma-3-12B	53.6	15.5	25.5	59.2	64.3	55.9
17-18	Laniqo	53.4	7.6	37.0	61.2	57.4	51.5
19-34	15 systems not human-evaluated		...				

English→Icelandic

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	**Human**	87.5		87.4	88.4	86.8	87.3
2-2	Gemini-2.5-Pro	77.6	1.8	79.8	68.8	83.1	77.9
3-4	Erlendur	68.3	2.2	69.4	61.2	72.0	71.2
3-4	GPT-4.1	68.0	1.9	74.7	67.4	63.6	69.1
5-5	Shy-hunyuan-MT	63.2	1.0	51.3	67.0	66.6	65.2
6-6	TowerPlus-9B[M]	57.4	3.9	46.0	57.1	65.5	56.3
7-7	ONLINE-B	51.8	4.4	43.4	45.6	59.3	57.3
8-10	Claude-4❓	47.8	5.2	43.0	48.9	45.4	56.3
8-10	TowerPlus-72B[M]	46.3	5.7	39.5	39.2	52.1	54.5
8-10	TranssionTranslate❓	46.2	5.8	29.1	45.5	52.6	55.6
11-11	AMI	39.9	7.4	47.8	35.5	39.8	37.5
12-12	GemTrans	34.8	7.0	25.0	32.4	39.5	41.4
13-14	SalamandraTA	31.3	8.6	28.0	23.9	33.9	41.6
13-15	UvA-MT	30.6	6.8	23.8	23.7	37.3	36.9
14-15	CommandA-WMT🚫	29.0	6.8	9.6	36.0	31.6	36.5
16-16	NLLB	24.1	15.2	22.8	21.1	25.1	28.2
17-17	IRB-MT	20.7	11.9	6.2	21.2	24.7	29.5
18-18	Gemma-3-12B	16.5	13.8	8.4	12.8	19.1	26.6
19-19	Llama-3.1-8B🚫	10.5	24.9	10.5	4.4	13.4	14.4
20-35	15 systems not human-evaluated		...				

English→Serbian (Cyrilic)

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Gemini-2.5-Pro	94.2	3.0	97.3	92.8	96.0	89.2
2-3	GPT-4.1	92.5	3.4	98.6	90.5	91.9	89.5
2-4	Shy-hunyuan-MT	92.2	1.0	94.4	90.0	94.3	88.8
3-4	ONLINE-B	90.6	6.1	97.7	90.9	90.6	81.1
5-5	Claude-4❓	90.0	6.8	96.1	86.4	93.2	81.8
6-6	**Human**	88.7		83.8	93.5	88.4	86.9
7-7	TranssionTranslate❓	85.1	8.0	88.7	87.7	86.6	73.2
8-9	GemTrans	81.5	4.6	88.3	78.8	79.7	81.6
8-9	DeepSeek-V3❓	78.7	8.6	89.8	87.0	61.7	84.5
10-11	IRB-MT	77.6	9.9	81.7	80.4	77.3	68.4
10-15	DLUT_GTCOM	77.2	9.3	72.3	80.0	78.4	75.9
11-14	CommandA-WMT🚫	76.5	7.0	59.1	78.8	84.8	77.9
10-15	UvA-MT	76.2	5.8	63.4	77.9	83.0	75.4
11-15	SalamandraTA	75.5	8.8	62.0	82.8	78.2	73.9
13-15	Gemma-3-12B	74.8	12.1	70.1	71.6	81.9	72.2
16-17	CUNI-SFT	60.9	13.5	65.6	58.5	61.8	57.4
16-17	Llama-3.1-8B🚫	58.4	19.4	53.7	63.6	60.8	50.3
18-18	NLLB	53.5	19.8	41.6	60.5	59.3	44.1
19-19	EuroLLM-9B[M]🚫	41.8	22.3	73.4	30.3	41.6	22.9
20-34	14 systems not human-evaluated		...				

Czech→German

Rank	System	Human	AutoRank	dialogue	edu	news	social	speech
1-1	Gemini-2.5-Pro	90.7	2.5	94.5	92.9	86.2	94.0	89.8
2-4	GPT-4.1	89.5	2.4	90.2	90.2	86.6	95.7	85.9
2-4	Claude-4	88.8	4.8	92.1	88.4	86.5	93.1	85.8
2-6	DeepSeek-V3❓	88.1	3.5	92.3	91.5	85.0	93.2	81.0
4-7	Shy-hunyuan-MT	87.2	1.0	89.9	81.7	87.6	92.4	84.7
4-8	Mistral-Medium	87.0	4.2	91.8	86.6	87.0	90.1	80.7
5-7	CommandA	86.8	4.8	91.0	83.1	85.1	93.6	83.1
8-8	CommandA-WMT	85.6	2.1	87.6	86.0	83.8	91.6	80.0
9-12	**Human**	82.8		93.6	88.1	75.5	81.1	84.1
9-13	GemTrans	82.6	6.3	87.1	83.6	79.9	87.7	77.3
9-13	Gemma-3-27B	82.0	7.2	86.7	85.4	74.6	87.9	80.8
9-13	Wenyiil	82.0	10.9	88.2	72.3	86.9	86.6	74.9
10-15	Algharb	80.9	13.2	90.5	72.0	88.2	81.6	71.1
13-15	TowerPlus-9B[M]	79.8	10.3	81.2	81.5	74.9	89.6	73.9
13-15	UvA-MT	79.5	7.0	94.6	69.0	73.0	89.9	79.8
16-19	CUNI-MH-v2	77.2	14.2	77.1	73.0	73.8	87.9	75.6
16-18	Gemma-3-12B	76.8	11.5	76.2	69.0	75.5	89.0	74.2
16-18	SRPOL	76.7	11.0	79.7	69.1	73.8	90.8	71.9
19-19	Yolu	75.3	9.3	91.5	63.3	71.3	85.2	72.9
20-21	IRB-MT	71.7	12.4	63.0	70.9	65.2	86.5	72.3
20-21	Laniqo	70.0	10.3	76.3	70.0	66.7	74.4	66.0
22-42	20 systems not human-evaluated		...					

English→Czech

Rank	System	Human	AutoRank	dialogue	literary	news	social	speech
1-1	Gemini-2.5-Pro	88.7	3.4	91.4	96.1	86.5	84.4	87.6
2-2	Shy-hunyuan-MT	87.1	1.0	88.7	94.1	89.8	81.6	80.7
3-4	DeepSeek-V3❓	85.1	5.1	91.0	90.4	85.6	84.0	75.0
3-4	**Human**	84.5		86.4	88.3	84.0	84.1	80.0
5-6	CommandA-WMT	82.6	3.6	90.1	83.5	84.1	82.7	72.8
5-6	Wenyiil	82.4	4.5	82.9	81.2	83.6	82.8	81.1
7-9	GPT-4.1	80.8	4.0	91.3	70.6	80.7	84.2	81.0
7-9	Mistral-Medium❓	80.4	7.1	86.6	88.1	78.7	77.4	74.0
7-10	Claude-4❓	79.6	9.0	86.5	85.5	78.9	75.0	75.8
9-11	UvA-MT	78.6	6.5	85.6	86.4	70.6	84.2	68.7
10-14	Algharb	76.7	6.4	85.1	50.7	84.9	81.9	81.4
11-14	CommandA	76.4	8.8	88.1	75.6	77.9	73.2	71.4
11-15	Yolu	75.6	6.3	82.3	83.3	73.1	76.0	64.8
11-15	Gemma-3-27B	75.6	9.2	82.9	85.1	72.3	72.8	68.3
13-15	GemTrans	73.2	5.1	87.5	55.3	79.1	75.6	72.0
16-16	CUNI-MH-v2	71.0	12.1	75.7	77.4	76.1	65.7	58.8
17-18	SRPOL	67.5	8.7	74.9	67.7	75.3	58.9	61.5
17-19	Laniqo	66.1	8.8	51.1	79.6	67.7	64.3	59.1
18-19	TowerPlus-9B[M]	65.8	10.4	74.4	58.4	70.6	66.5	59.4
20-20	SalamandraTA	60.3	10.5	57.0	62.0	70.0	52.5	55.7
21-44	23 systems not human-evaluated		...					

English→Chinese

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	Algharb	88.4	4.2	95.0	87.7	88.4	81.9
2-4	Shy-hunyuan-MT	88.2	1.0	93.2	84.5	92.4	80.1
2-5	Claude-4	86.9	7.2	98.2	86.3	84.0	79.7
2-5	Wenyiil	86.3	4.0	89.5	80.3	91.4	82.0
3-6	DeepSeek-V3	85.0	7.3	94.5	83.8	82.5	80.1
5-10	GemTrans	84.4	5.0	94.2	80.7	85.3	76.7
6-11	Qwen3-235B	84.0	4.9	88.2	85.5	85.5	74.3
5-10	GPT-4.1	84.0	4.7	98.3	80.9	79.5	80.2
6-11	Gemini-2.5-Pro	83.8	4.0	82.1	83.2	85.5	83.7
5-10	UvA-MT	83.4	6.4	96.7	78.0	84.3	74.4
11-13	**Human**	82.1		92.8	74.2	83.7	78.3
11-15	CommandA-WMT	81.3	5.7	82.0	86.8	80.0	75.0
11-15	Llama-4-Maverick	80.7	8.1	83.9	81.1	82.3	73.5
12-16	Mistral-Medium ?	79.9	5.0	78.0	83.2	78.7	79.7
12-16	Yolu	79.0	4.9	84.5	82.9	76.9	71.1
14-17	SRPOL ⃠	77.7	10.5	68.8	79.4	85.7	70.8
16-18	IRB-MT	76.5	9.5	90.3	70.6	77.5	67.4
17-18	RuZh ?	75.7	10.6	84.1	73.2	77.1	66.9
19-19	Laniqo	70.5	9.3	83.0	72.4	63.6	65.7
20-40	20 systems not human-evaluated		...				

English→Japanese

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	**Human**	89.2		94.5	85.2	92.1	84.2
2-4	Gemini-2.5-Pro	85.8	2.5	87.3	82.5	87.7	86.0
2-6	Algharb	85.7	3.3	84.3	88.9	83.8	85.6
2-5	Mistral-Medium ?	84.8	5.5	98.4	77.1	83.3	82.6
3-6	Wenyiil	84.4	3.0	88.6	80.9	85.6	82.5
5-6	GPT-4.1	83.7	2.9	95.4	77.0	80.7	84.9
7-7	CommandA-WMT	82.2	3.7	83.3	85.2	78.0	83.1
8-12	Shy-hunyuan-MT	79.9	1.0	75.6	78.2	81.8	84.3
8-13	DeepSeek-V3 ?	79.3	4.7	82.9	80.0	74.1	82.7
8-13	Claude-4	79.3	5.8	86.5	76.1	72.8	86.3
8-13	UvA-MT	79.3	6.5	74.9	79.7	81.7	80.1
8-14	ONLINE-B	78.0	6.3	82.5	78.1	76.3	75.4
9-16	In2x ?	77.8	2.3	60.8	83.6	81.9	82.7
12-16	GemTrans	76.2	5.6	81.0	66.9	80.9	76.8
13-16	KIKIS	76.2	3.2	66.6	78.5	79.2	79.1
13-16	Systran	75.6	7.5	69.2	84.5	75.9	69.5
17-18	NTTSU	73.3	8.1	75.3	77.9	71.9	66.5
17-18	Yolu	72.6	6.1	72.0	76.4	70.7	71.0
19-19	Laniqo	67.8	9.5	49.0	72.0	81.4	61.6
20-45	25 systems not human-evaluated		...				

Czech→Ukrainian

Rank	System	Human	AutoRank	edu	news	social	speech
1-2	Gemini-2.5-Pro	92.9	1.1	96.8	93.4	91.6	89.4
1-3	GPT-4.1	92.1	1.3	94.0	92.3	92.9	88.9
2-3	Shy-hunyuan-MT	91.8	1.0	91.7	94.7	90.1	89.0
4-8	GemTrans	90.2	4.4	92.9	91.0	89.5	86.8
4-6	**Human**	90.1		93.0	92.6	85.5	88.0
4-10	Mistral-Medium ?	89.4	4.2	91.1	91.7	88.7	84.6
6-10	Claude-4 ?	89.1	3.7	91.4	92.4	88.7	81.3
4-10	DeepSeek-V3 ?	89.0	3.2	90.7	91.0	88.2	84.8
6-10	CommandA-WMT	88.7	1.3	87.3	89.6	91.2	85.7
6-10	Gemma-3-27B	88.6	5.0	89.1	91.3	88.5	83.7
11-12	CommandA	86.4	4.6	86.1	86.6	89.6	83.0
11-13	Wenyiil	85.7	5.4	72.9	93.6	89.6	81.3
12-15	TowerPlus-9B[M]	85.3	7.9	85.0	87.9	88.1	78.2
13-16	Algharb	84.1	7.2	74.7	93.8	87.2	73.9
13-17	UvA-MT	83.5	5.1	75.3	86.0	87.4	83.5
14-17	Laniqo	83.4	7.7	79.6	89.3	84.7	75.7
15-17	IRB-MT	82.7	9.1	77.2	86.7	84.4	79.5
18-19	SRPOL	80.8	7.8	74.3	88.4	80.9	74.6
18-19	Yolu	80.1	6.0	66.4	88.6	82.6	77.2
20-44	24 systems not human-evaluated		...				

Japanese→Chinese

Rank	System	Human	AutoRank	literary	news	social	speech
1-1	**Human**	-3.5		-3.6	-3.8	-3.0	-3.3
2-2	Gemini-2.5-Pro	-4.4	3.3	-3.9	-5.1	-2.2	-6.8
3-6	Algharb	-5.8	4.3	-6.5	-4.6	-5.1	-7.5
3-7	Claude-4	-5.9	6.4	-4.6	-4.6	-5.3	-11.5
3-7	Shy-hunyuan-MT	-6.1	1.0	-5.2	-5.4	-4.5	-11.1
3-7	GPT-4.1	-6.2	4.5	-4.5	-7.1	-4.7	-9.9
4-7	Wenyiil	-6.9	4.5	-6.4	-6.5	-5.4	-10.5
8-10	CommandA-WMT	-7.7	5.2	-7.1	-6.3	-4.5	-15.7
8-10	DeepSeek-V3	-8.1	6.5	-8.9	-5.9	-4.0	-16.3
8-13	Kaze-MT	-8.6	3.9	-8.1	-8.4	-6.0	-13.1
10-13	Mistral-Medium 🚫	-10.0	6.6	-12.2	-7.3	-6.4	-15.8
10-13	In2x 🚫	-10.0	3.0	-9.2	-10.4	-7.9	-13.8
10-13	Qwen3-235B	-10.9	7.6	-14.3	-7.5	-5.7	-17.9
14-15	GemTrans	-10.9	6.6	-11.0	-9.1	-8.4	-17.5
14-15	NTTSU	-11.3	5.9	-10.5	-9.4	-6.3	-22.8
16-17	Yolu	-12.6	7.1	-14.2	-7.6	-9.1	-23.8
16-17	TowerPlus-9B[M]	-13.3	11.5	-12.4	-9.4	-8.2	-29.3
18-18	IRB-MT	-13.9	12.4	-16.2	-10.8	-11.6	-18.9
19-19	Laniqo	-18.3	11.3	-20.4	-14.6	-14.6	-26.3
20-42	22 systems not human-evaluated		...				

English→Korean

Rank	System	Human	AutoRank	literary	news	social	speech
1-3	**Human**	-1.9		-2.4	-1.7	-1.7	-1.4
1-3	Shy-hunyuan-MT	-2.5	1.0	-3.0	-2.2	-1.0	-2.4
1-3	Gemini-2.5-Pro	-2.7	2.5	-3.5	-3.8	-0.7	-1.5
4-6	GPT-4.1	-3.3	2.9	-4.2	-3.7	-1.6	-2.1
4-7	Claude-4	-3.4	4.4	-3.1	-5.8	-2.2	-2.7
4-7	DeepSeek-V3 🚫	-3.8	5.1	-4.1	-4.5	-3.2	-2.8
5-10	GemTrans	-4.1	5.0	-4.1	-8.0	-1.7	-2.2
7-12	CommandA-WMT	-4.3	2.9	-4.3	-5.5	-0.7	-4.7
5-12	Wenyiil	-4.3	3.0	-6.2	-4.5	-1.1	-2.4
5-12	Algharb	-4.4	3.1	-6.3	-5.1	-1.6	-1.6
8-15	Mistral-Medium 🚫	-4.7	6.1	-5.6	-6.1	-1.8	-3.2
7-15	CommandA	-4.7	6.0	-3.9	-7.6	-2.2	-4.9
11-16	UvA-MT	-5.2	4.3	-5.5	-8.4	-1.2	-3.7
11-16	Qwen3-235B	-5.5	6.5	-6.3	-7.2	-1.9	-4.2
11-16	IRB-MT	-5.6	8.6	-6.3	-8.1	-3.2	-3.5
13-16	Gemma-3-12B	-5.9	9.2	-5.9	-8.4	-2.9	-4.9
17-18	TowerPlus-9B[M]	-7.2	10.1	-7.4	-8.2	-2.9	-7.8
17-18	Yolu	-7.3	7.0	-7.3	-11.3	-2.3	-6.3
19-19	Laniqo	-9.1	9.2	-10.6	-12.6	-3.6	-5.9
20-37	17 systems not human-evaluated		...				

E Analysis of Human Evaluation Scores

Figure 4 shows the correlation between ranks obtained from human evaluation ranks and automatic evaluation (AUTORANK) for each system. Figure 5 shows the distribution of human evaluation ranks across all systems. Finally, Figure 6 and Figure 7 break down the distribution of average human evaluation scores by language pair and by domain, respectively.

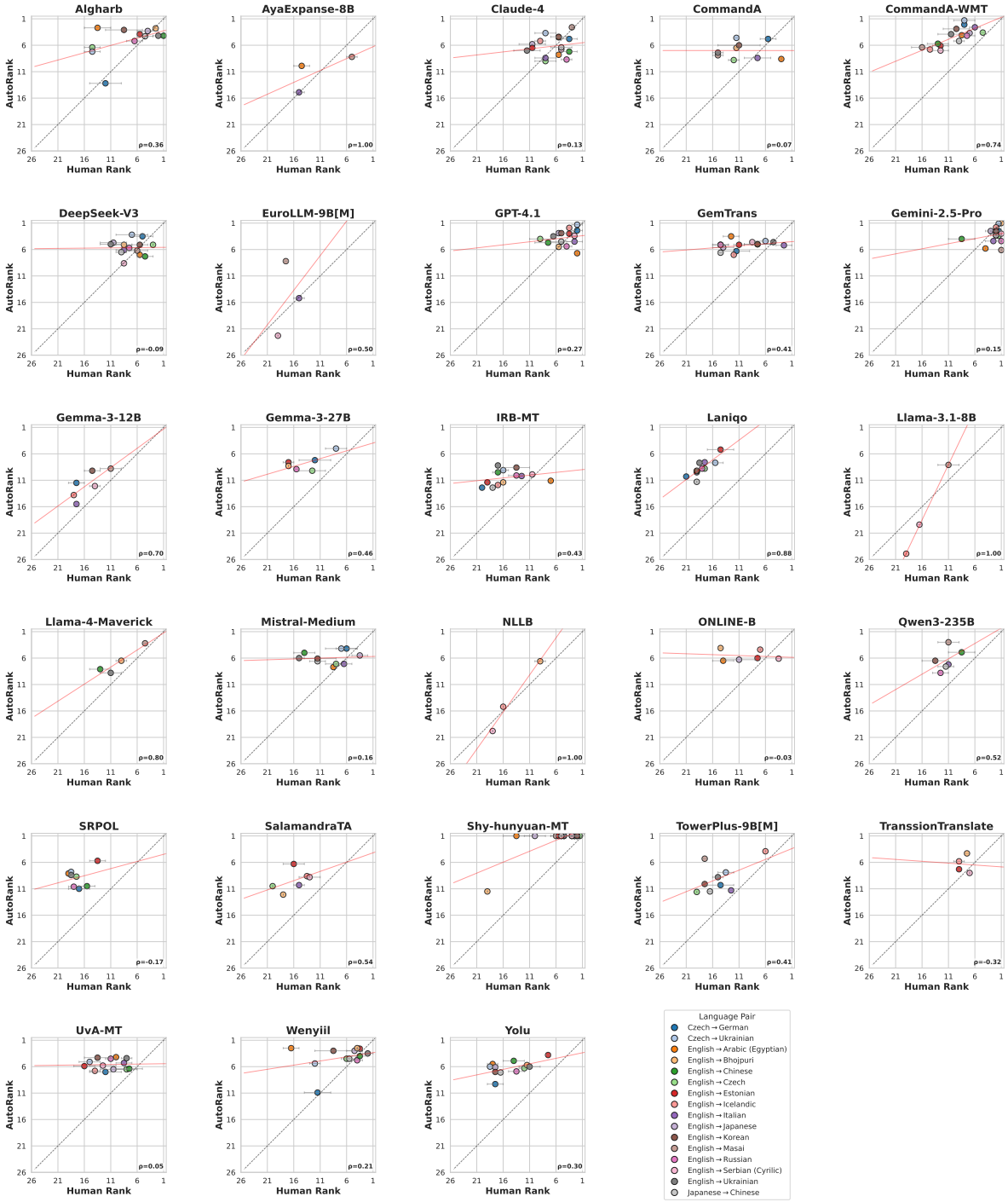


Figure 4: Correlation between automatic (AUTORANK) and human evaluation ranks by model (lower=better). Whiskers indicate the range of human ranks. The gray diagonal represents perfect correlation; points above this line mean AUTORANK ranked a model higher than humans, and vice versa. Colors denote language pairs, and the red line shows the Spearman correlation (ρ).

Figure 5: Distribution of ranks from human evaluation for each system, with whiskers indicating the assigned ranges. Systems are sorted by the number of "wins" (which refers to the situation when a system is being ranked first or has a rank range that includes the first position) and then by average rank.

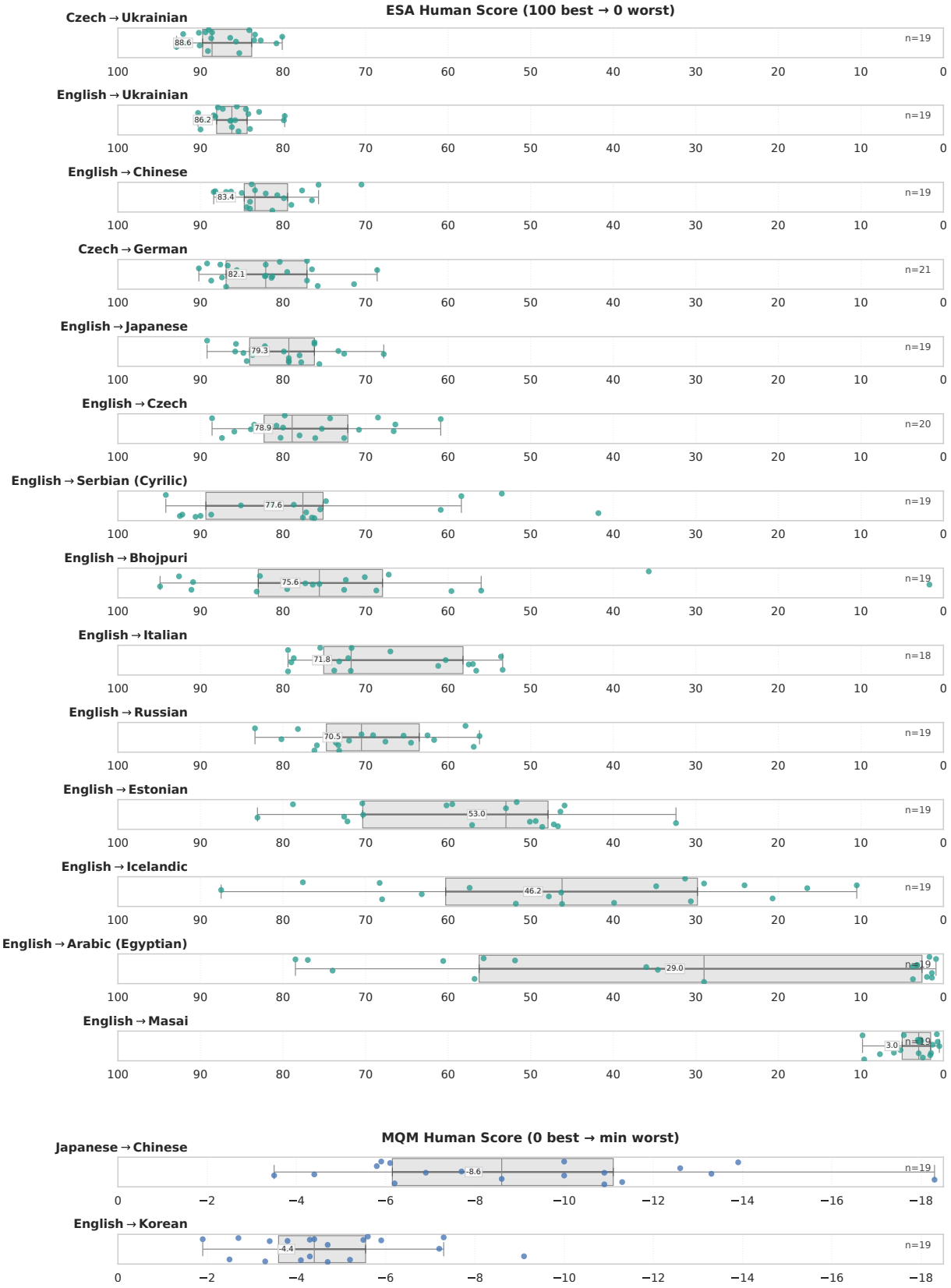


Figure 6: The distribution of human evaluation scores for each language pair. Pairs are grouped by their evaluation protocol, with ESA at the top and MQM at the bottom.

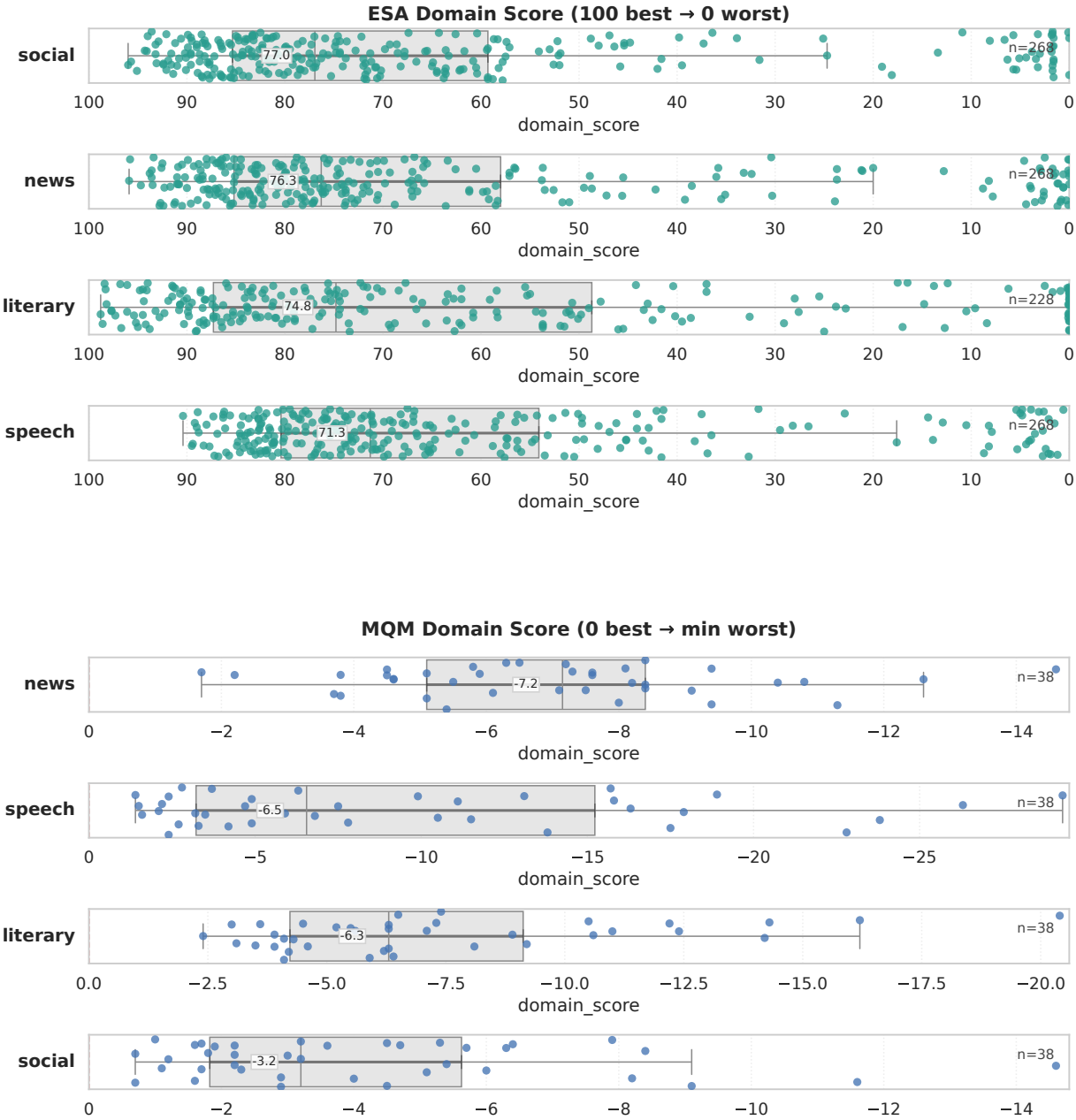


Figure 7: Distribution of scores by domain for four main domains. ESA scores are presented at the top, while MQM scores are presented at the bottom.

F Dataset Statistics

Statistics for the parallel training data provided for the shared task are shown in Tables 18 and 19.

Dataset	Segments	Tokens		Characters	
		Source	Target	Source	Target
Bhojpuri→English	Segs	Bhojpuri	English	Bhojpuri	English
OPUS	2.43M	23.07M	19.10M	270.18M	105.16M
Czech→German	Segs	Czech	German	Czech	German
OPUS	136.54M	1.47B	1.61B	10.65B	11.42B
LinguaTools-wikititles-2014	2.39M	4.65M	4.28M	40.00M	40.23M
Tilde	2.04M	36.30M	38.02M	288.88M	307.45M
Facebook-wikimatrix-1	1.60M	20.74M	22.47M	151.14M	162.61M
Statmt-news_commentary-18.1	244.83k	4.82M	5.45M	37.02M	41.07M
(Total)	142.82M	1.54B	1.68B	11.16B	11.97B
Czech→Ukrainian	Segs	Czech	Ukrainian	Czech	Ukrainian
OPUS	17.15M	138.66M	137.78M	0.97B	1.65B
Facebook-wikimatrix-1	848.96k	10.43M	10.07M	75.97M	127.31M
ELRC	130.00k	2.48M	2.56M	19.61M	35.26M
(Total)	18.13M	151.57M	150.41M	1.07B	1.81B
English→Arabic	Segs	English	Arabic	English	Arabic
OPUS	304.22M	4.65B	4.21B	28.48B	44.10B
Statmt-ccaligned-1	25.31M	355.78M	343.52M	2.27B	3.58B
LinguaTools-wikititles-2014	4.82M	11.15M	10.91M	84.51M	129.17M
Facebook-wikimatrix-1	1.97M	38.55M	35.77M	242.74M	376.25M
Statmt-tedtalks-2_clean	341.89k	6.17M	5.41M	34.54M	54.49M
Statmt-news_commentary-18.1	193.67k	8.94M	11.70M	57.33M	127.15M
(Total)	336.86M	5.07B	4.61B	31.17B	48.37B
English→Czech	Segs	English	Czech	English	Czech
OPUS	237.54M	2.85B	2.48B	17.02B	17.82B
ParaCrawl-paracrawl-9	50.63M	692.12M	626.34M	4.33B	4.68B
Statmt-ccaligned-1	12.73M	148.71M	135.81M	936.99M	1.01B
LinguaTools-wikititles-2014	4.81M	11.36M	9.67M	83.77M	81.29M
Facebook-wikimatrix-1	2.09M	33.56M	29.66M	206.82M	216.62M
Tilde	2.09M	42.26M	38.26M	276.52M	303.75M
ELRC	1.96M	37.18M	33.00M	243.79M	262.52M
EU	1.92M	34.27M	30.09M	222.84M	232.92M
Statmt-europarl-10	644.43k	15.63M	13.00M	94.31M	98.14M
Statmt-wikititles-3	410.94k	1.03M	965.62k	7.47M	7.57M
Statmt-news_commentary-18.1	265.37k	5.71M	5.19M	36.22M	39.81M
Statmt-commoncrawl_wmt13-1	161.84k	3.35M	2.93M	20.66M	20.75M
Neulab-tedtalks_train-1	103.09k	2.10M	1.77M	10.58M	10.39M
(Total)	315.37M	3.88B	3.40B	23.49B	24.78B
English→Estonian	Segs	English	Estonian	English	Estonian
OPUS	121.36M	1.83B	1.38B	11.18B	11.02B
ELRC	9.09M	201.49M	144.73M	1.29B	1.25B
ParaCrawl-paracrawl-9	8.54M	136.60M	103.32M	846.64M	840.74M
Statmt-ccaligned-1	4.11M	54.21M	43.28M	339.16M	338.17M
Tilde	2.06M	41.65M	30.28M	272.67M	271.35M
EU	2.03M	36.68M	26.85M	237.87M	231.57M
Facebook-wikimatrix-1	955.55k	15.41M	11.78M	96.18M	95.33M
Statmt-europarl-7	649.59k	15.68M	11.21M	94.64M	91.44M
Neulab-tedtalks_train-1	10.74k	215.97k	171.65k	1.09M	1.04M
(Total)	148.81M	2.33B	1.76B	14.36B	14.14B
English→Icelandic	Segs	English	Icelandic	English	Icelandic
OPUS	24.26M	292.15M	274.41M	1.70B	1.84B
ParaCrawl-paracrawl-9	2.97M	45.10M	42.66M	266.09M	292.17M
ParIce-eea_train-20.05	1.70M	26.75M	24.24M	170.36M	179.49M
Statmt-ccaligned-1	1.19M	18.63M	17.80M	115.58M	124.36M
Tilde	420.71k	6.31M	6.10M	41.71M	45.26M
ParIce-ema_train-20.05	399.09k	6.13M	5.94M	40.41M	43.90M
Facebook-wikimatrix-1	313.88k	5.66M	4.77M	34.53M	34.04M
Statmt-wikititles-3	50.18k	98.99k	88.35k	722.24k	763.33k
EU	4.72k	54.43k	52.31k	369.04k	398.50k
(Total)	31.31M	400.87M	376.06M	2.37B	2.56B

Table 18: Statistics for parallel training data provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

Dataset	Segments	Tokens		Characters	
		Source	Target	Source	Target
English→Korean	Segs	English	Korean	English	Korean
OPUS	138.12M	1.64B	1.31B	9.84B	12.28B
Statmt-ccaligned-1	9.03M	98.69M	84.80M	635.05M	744.99M
LinguaTools-wikititles-2014	4.83M	11.62M	9.32M	84.86M	90.51M
ParaCrawl-paracrawl-1_bonus	4.00M	61.96M	48.70M	371.75M	433.95M
Facebook-wikimatrix-1	1.35M	21.63M	15.66M	135.00M	161.17M
Neulab-tedtalks_train-1	205.64k	4.29M	2.97M	21.55M	26.31M
ELRC	3.27k	67.72k	45.95k	424.80k	471.77k
(Total)	157.54M	1.84B	1.48B	11.09B	13.74B
English→Russian	Segs	English	Russian	English	Russian
OPUS	479.12M	7.32B	6.39B	44.88B	83.67B
Statmt-ccaligned-1	69.26M	0.97B	864.09M	6.18B	11.32B
Statmt-backtrans_ruen-wmt20	39.36M	746.47M	596.28M	4.47B	7.75B
LinguaTools-wikititles-2014	13.57M	33.05M	28.99M	245.88M	421.65M
ParaCrawl-paracrawl-1_bonus	5.38M	101.31M	80.41M	632.54M	1.06B
Facebook-wikimatrix-1	5.20M	86.79M	76.48M	537.73M	0.97B
Statmt-wikititles-3	1.19M	3.13M	2.88M	22.80M	39.34M
Statmt-yandex-wmt22	1.00M	21.25M	18.68M	130.99M	250.76M
Statmt-commoncrawl_wmt13-1	878.39k	18.77M	17.40M	116.16M	214.59M
Statmt-news_commentary-18.1	377.66k	8.72M	8.11M	55.68M	112.13M
Neulab-tedtalks_train-1	208.46k	4.37M	3.69M	21.96M	36.77M
ELRC	39.50k	891.98k	792.00k	5.73M	10.87M
Tilde	34.27k	752.66k	702.81k	4.83M	9.97M
(Total)	615.62M	9.31B	8.09B	57.31B	105.86B
English→Serbian	Segs	English	Serbian	English	Serbian
OPUS	127.45M	1.33B	1.17B	7.57B	9.99B
Statmt-ccaligned-1	1.99M	38.73M	34.34M	235.07M	399.09M
Facebook-wikimatrix-1	1.21M	20.95M	18.81M	129.91M	209.19M
Neulab-tedtalks_train-1	136.90k	2.79M	2.38M	14.05M	14.40M
Tilde	2.02k	46.81k	45.16k	303.95k	491.17k
ELRC	856	14.50k	13.28k	93.28k	149.56k
(Total)	130.79M	1.39B	1.22B	7.95B	10.62B
English→Ukrainian	Segs	English	Ukrainian	English	Ukrainian
OPUS	151.87M	2.68B	2.33B	16.50B	29.37B
ParaCrawl-paracrawl-1_bonus	13.35M	505.83M	487.47M	3.28B	6.04B
Statmt-ccaligned-1	8.55M	119.38M	104.10M	755.38M	1.33B
Facebook-wikimatrix-1	2.58M	41.55M	35.59M	257.56M	447.33M
ELRC	1.16M	16.65M	13.15M	110.37M	194.76M
Neulab-tedtalks_train-1	108.50k	2.25M	1.94M	11.33M	18.45M
Tilde	1.63k	36.07k	34.18k	237.96k	477.91k
(Total)	177.62M	3.36B	2.97B	20.92B	37.40B
English→Chinese	Segs	English	Chinese	English	Chinese
OPUS	221.88M	3.25B	392.85M	19.99B	17.76B
Statmt-backtrans_enzh-wmt20	19.76M	364.22M	32.72M	2.16B	1.96B
Statmt-ccaligned-1	15.18M	155.93M	42.42M	1.04B	1.13B
ParaCrawl-paracrawl-1_bonus	14.17M	217.60M	46.40M	1.34B	1.18B
LinguaTools-wikititles-2014	6.66M	16.16M	7.79M	118.50M	112.12M
Facebook-wikimatrix-1	2.60M	49.87M	5.00M	311.07M	277.84M
Statmt-wikititles-3	921.96k	2.37M	973.44k	17.82M	16.28M
Statmt-news_commentary-18.1	442.93k	9.80M	799.74k	62.67M	55.16M
Neulab-tedtalks_train-1	5.54k	95.63k	23.52k	476.98k	399.81k
ELRC	2.98k	91.23k	7.36k	591.36k	644.17k
(Total)	281.63M	4.07B	528.99M	25.05B	22.49B
Japanese→Chinese	Segs	Japanese	Chinese	Japanese	Chinese
OPUS	19.74M	46.43M	46.87M	1.44B	1.08B
KECL-paracrawl-2wmt24	4.60M	27.88M	29.51M	0.97B	704.98M
LinguaTools-wikititles-2014	1.66M	1.97M	1.97M	35.18M	27.48M
Facebook-wikimatrix-1	1.33M	2.36M	2.12M	145.10M	113.60M
KECL-paracrawl-2	83.89k	552.50k	633.77k	18.86M	14.11M
Neulab-tedtalks_train-1	5.16k	19.57k	22.30k	490.89k	375.98k
Statmt-news_commentary-18.1	1.62k	2.59k	2.17k	272.83k	197.25k
(Total)	27.42M	79.23M	81.13M	2.61B	1.94B

Table 19: Statistics for parallel training data provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

G Analysis of English→Serbian Outputs

For the English→Serbian language direction, we tested generation of translations in both Cyrillic and Latin scripts, and we can therefore compare the use of the two scripts for each system. Table 20 shows the amount of overlap between generation in the Latin and Cyrillic scripts, measured using the word bigram F1-score (w2F).

SYSTEM	W2F
ONLINE-B (c)	99.8
ONLINE-G (c)	98.0
GemTrans	92.2
UvA-MT (l)	82.6
Claude-4	79.2
GPT-4.1	75.2
Gemini-2.5-Pro	73.7
CUNI-SFT (l)	73.5
Llama-3.1-8B (l)	73.0
EuroLLM-22B	73.0
Gemma-3-12B	71.5
Gemma-3-27B	68.0
Llama-4-Maverick	66.8
DeepSeek-V3	66.5
IRB-MT	65.7
AyaExpanse-8B	62.6
TowerPlus-9B	60.8
Qwen2.5-7B	60.2
CommandA	59.9
Shy	58.5
SalamandraTA	57.9
TranssionTranslate	57.7
Qwen3-235B	54.9
IR-MultiagentMT	54.4
CommandR7B	50.3
Mistral-7B	47.2
TowerPlus-72B	47.1
EuroLLM-9B	46.5
AyaExpanse-32B	41.2

Table 20: Content overlap between Cyrillic and Latin script translations for English→Serbian, measured with the word bigram F1-score (w2F).

Findings of the WMT25 Multilingual Instruction Shared Task: Persistent Hurdles in Reasoning, Generation, and Evaluation

Tom Kocmi Cohere
Ekaterina Artemova Toloka AI
Eleftherios Avramidis DFKI
Eleftheria Briakou Google
Pinzhen Chen University of Edinburgh
Marzieh Fadaee Cohere Labs
Markus Freitag Google
Roman Grundkiewicz Microsoft
Yupeng Hou UC San Diego
Philipp Koehn JHU
Julia Kreutzer Cohere Labs
Saab Mansour Amazon
Stefano Perrella Sapienza University
Lorenzo Proietti Sapienza University
Parker Riley Google
Eduardo Sánchez Meta
Patrícia Schmidtová Charles University
Mariya Shmatova Toloka AI
Vilém Zouhar ETH Zurich

Abstract

The WMT25 Multilingual Instruction Shared Task (MIST) introduces a benchmark to evaluate large language models (LLMs) across 30 languages. The benchmark covers five types of problems: machine translation, linguistic reasoning, open-ended generation, cross-lingual summarization, and LLM-as-a-judge. We provide automatic evaluation and collect human annotations, which highlight the limitations of automatic evaluation and allow further research into metric meta-evaluation. We run on our benchmark a diverse set of open- and closed-weight LLMs, providing a broad assessment of the multilingual capabilities of current LLMs. Results highlight substantial variation across sub-tasks and languages, revealing persistent challenges in reasoning, cross-lingual generation, and evaluation reliability. This work establishes a standardized framework for measuring future progress in multilingual LLM development.

1 Introduction

We are witnessing rapid development of multilingual large language models (LLMs). However, as pointed out by recent works (Kreutzer et al., 2025; Wu et al., 2025; Cruz Blandón et al., 2025), multilingual benchmarks lack comprehensiveness, scientific rigor, and consistent adoption across research labs, undermining their value in guiding multilingual LLM development. Among common problems are benchmark contamination (Ahuja et al., 2024), label noise (Chalamalasetti et al., 2025), reliance on non-native (machine-)translated instances (Chen et al., 2024b), and inconsistent evaluation pipelines. For instance, some leading LLM descriptions report multilinguality solely through translated MMLU. There is a mist surrounding multilingual evaluation that we aim to see through with this year's MIST shared task.

We introduce a novel multilingual evaluation benchmark that systematically assesses several key capabilities of LLMs across 30 diverse languages using the following sub-tasks:

- **Machine Translation (MT)**: A standardized, well-defined cross-lingual task.
- **Linguistic Reasoning (LR)**: Structured linguistic problem solving in multiple languages.
- **Open-Ended Generation (OEG)**: Using localized open-ended questions to assess language proficiency instead of specific capabilities.
- **Cross-lingual Summarization (XLSum)**: Synthesizing multilingual content from multiple documents written in different languages.
- **LLM-as-a-Judge**: Testing the effectiveness of LLMs in evaluating the quality of outputs in other sub-tasks that do not have definitive answers (MT, OEG, and XLSum).

We benchmark several of the most commonly used open- and closed-weight systems on our tests. These tests provide a multi-faceted evaluation framework that highlights the strengths and limi-

Proceedings of the Tenth Conference on Machine Translation, pages 414–435
November 8-9, 2025

Linguistic Reasoning	Here are some word combinations in Hadza and their English translations: 1. chutisa zzokwanako: the giraffe's neck 2. athuitcha slimibii: the men's axe (for collecting honey) [. . .] Translate into Hadza: the male impalas' horns
Open-Ended Generation	As a news reporter, write an article about the opening of a new shopping complex, including who will enjoy it and what activities are available.
Cross-lingual Summarization	Fass bitte diese 6 Bewertungen eines Produkts auf Amazon auf Deutsch zusammen. Fleetwood Mack är som de är. Sköna att lyssna på. I am super pleased with my purchase and would order from this seller again. [...]
Machine Translation	You are a professional Czech-to-Ukrainian translator, tasked with providing translations for use in Ukraine. [...]
LLM-as-a-judge	Score the response generated by a system to a user's request in Lithuanian on a Likert scale from 1 to 7. The quality levels associated with numerical scores are provided below: [...]

Table 1: Example prompts for each sub-task.

tations of current LLMs across diverse linguistic phenomena while drawing on the rigorous principles established within the MT evaluation research.

In addition to automatic metrics, we conduct human evaluation for all sub-tasks without definitive answers, which is then used to assess LLM-as-a-judge systems. Test sets, system outputs, and human judgments are released with a permissive license.[1]

2 Data and Methodology

In this section, we describe the datasets and preparation steps used for each of the sub-tasks in our benchmark. For every sub-task, we curated or adapted data across up to 30 languages. The high-level statistics are in Table 2. The following sections detail the sources of the data, the translation or localization processes applied, and any additional filtering or validation steps specific to each sub-task.

2.1 Linguistic Reasoning

The data for the linguistic reasoning sub-task were sourced from the 2024 International Linguistics Olympiad (IOL). In this olympiad, high school students compete in solving linguistic puzzles. Problems and solutions are released online and manually translated into the participants' languages. Previous benchmarks built from previous linguistics olympiads in English, such as Linguini (Sánchez et al., 2024) and LINGOLY (Bean et al., 2024), have shown that this type of puzzle is challenging even for the best LLMs. To evaluate multilinguality, we propose a multilingual version of this problem. This enables us to not only benchmark LLMs on challenging, unseen problems but also measure language disparities. The key to these puzzles is not retrieving acquired knowledge, but rather applying reasoning.

	Langs	Samples per lang
Machine Translation	30	384
Linguistic Reasoning	15	90
Open Ended Generation	20	100
Cross-lingual Summarization	14	350
LLM-as-a-judge MT	16	1520
LLM-as-a-judge OEG	10	2256
LLM-as-a-judge XLSum	14	3200

Table 2: Number of languages and the number of samples (prompts) for each language or language pair in case of MT.

From problem PDFs to evaluation prompts IOL problems and solutions are published as PDFs under the CC-BY-SA 4.0 license.[2] They are typeset in the same LaTeX, format for all languages, which motivates our approach of tuning an automatic extraction for English, and then transferring it to the other languages. First, we manually extract questions and solutions from the PDFs for the five tasks of IOL (languages are Koryak, Hadza, Komnzo, Dâw, and Yanyuwa), which includes breaking tasks into sub-tasks (e.g. turning a matching task with four phrases to match across languages into four individual tasks), capturing metadata such as task authors, unifying task formulations and formats across task types. Then, we prompt an LLM to repeat this process for the other languages.[3] Last, with the help of human annotators that are proficient in the respective languages, we fix any errors, post-edit for cross-task consistency and translate task-level instructions.[4] This yields a total of

[1] github.com/wmt-conference/wmt-mist

[2] ioling.org

[3] A detailed description of this process is in Appendix A.

[4] Early attempts to translate tasks automatically rather than relying on parsing the language-specific solutions, quickly showed that automatic translation is not well-equipped for this task (at least not out-of-the-box) because the task involves disambiguating many single-word terms without much context (e.g. the word " roast" could be translated as noun or verb and in a literal or abstract sense), and special handling of

Grounding	localized	46
	generic	54
Type	brainstorming	23
	creative	35
	informational	25
	professional	17
Available Locale	ar_EG, bn_BD, cs_CZ, de_DE, el_GR, en_US, fa_IR, hi_IN, is_IS, id_ID, it_IT, ja_JP, kn_IN, ko_KR, ro_RO, ru_RU, sr_RS (both Latin and Cyrillic), uk_UA, zh_CN.	

Table 3: A breakdown of the 100 test questions in the open-ended generation sub-task.

90 prompts per language, covering five task types (classification (4), editing (1), fill-in-blanks (20), mapping (24) and translation (41)) for 15 languages (Chinese, Czech, Dutch, English, Estonian, French, German, Japanese, Korean, Persian, Portuguese, Russian, Spanish, Swedish, Ukrainian). Languages were chosen based on overlap with the 30 languages from the General MT task. For the final evaluation prompts, we add a simple instruction for context and answer format to each puzzle (e.g. English: "Solve the following linguistic puzzle with the help of the given context. The last line of your response should only contain the solution within square brackets [], nothing else."). We chose not to explicitly prompt the models for reasoning in order to avoid introducing any reasoning instruction bias and favoring models that are explicitly trained for reasoning. As a result, we can analyze how much each model tends to reason about each of these, but without having any expectations on the correctness, form, or volume of reasoning traces.

2.2 Open-Ended Generation

In this sub-task, we test multilingual language proficiency, e.g. generating native-sounding, useful, and coherent responses. Below a language's surface form are culture, values, and knowledge, so we also want to test LLMs' true ability grounded in the use of each language. The core motivation behind this is that LLMs sound native in English, but their responses in other languages are non-natural, contain English phenomena, or sound robotic (Guo et al., 2025). While some open-ended generation test sets exist, e.g. mArenaHard (Cohere Labs et al., 2024), they are often translated from English (Chen et al., 2024b) and skewed towards narrow domains like coding and math, which are not typical multilingual LLM use cases. Therefore, we focus on building a test set that asks native open-ended questions in many different domains, rather than specific tasks, e.g. writing a news article about a topic.

We prepared 100 questions manually with the help of LLMs, localized them into different languages, and asked native speakers to post-edit them to make them more natural and native. As a result, this multilingual test set contains comparable questions localized into each locale (language and country/region). The details of the process for question creation and localization are as follows.

English question creation First, we obtained a set of 100 English questions via two complementary workflows:

1. Three authors of this paper wrote a small pool of diverse questions.
2. We iteratively fed five randomly selected human-authored questions to two LLMs (GPT-4.1-mini and Command A), asking for a new question.
3. Then we manually inspected and post-edited these questions while mixing them with the original human-written questions.

To ensure each question's applicability to multiple locales, all locale-specific mentions stay as placeholders, e.g., using "`{language}`" instead of "English" in prompts like "Please suggest an idiom in `{language}`".

Localization and quality control We localized the English questions into 19 more unique language-writing script combinations, each of which is designated a country too, to better ground questions in locales. A full list of locales is available in Table 3, and the five-step process is detailed below:

1. Localization: We used four LLMs[5] to localize the questions and replace placeholders with locale-specific content, yielding four candidate variants per question.
2. Baseline: We also generated a reference translation for each question using Google Translate.
3. Sanity Check: To prevent LLMs from answering the question rather than faithfully localizing it, using the Google Translate version as a reference, we discarded any model variant that has a chrF score below 30 or exceeds the baseline

grammatical annotations such as for singular and plural.

[5]DeepSeek V3, Gemini 2.5 Pro, Command A, GPT 4.1

length by more than 50% per NLLB-200 tokenization.

4. Selection: From the remaining variants, we discarded the lowest-scoring chrF candidate, then randomly selected a variant translation for inclusion. If no variant passed filtering, we defaulted to the Google Translate baseline.
5. Human Inspection: we conducted a review and applied post edits if necessary for all languages to minimize non-nativeness and translationese.

Nature of the questions In Table 3, we present a breakdown of the types of test questions and expected responses. By counting placeholders in the seed English questions, we find that 46 questions explicitly mention a language/country-specific entity (i.e., locale-grounded), and 54 questions are more generic. Using Gemini 2.5 Pro followed by human inspection, we classified the nature of the expected responses into one of "brainstorming", "creative", "informational", or "professional". It is worth noting that while we assigned only one label to each question, the labels are not strictly mutually exclusive.

2.3 Cross-lingual Summarization

Our cross-lingual summarization dataset combines multilingual review data from two complementary sources: Amazon product reviews and Google Maps restaurant reviews. The dataset construction process involved systematic sampling, language balancing, and content filtering to ensure high-quality cross-lingual evaluation data.

Data collection We integrated data from two distinct domains to maximize linguistic diversity: Amazon product reviews for consumer products, and Google Maps restaurant reviews for restaurants This resulted in an initial scraped dataset of 12,040 reviews spanning 853 products and restaurants. Each data item was paired with a product or restaurant-specific summarization prompt in 14 target languages: Arabic (Egyptian), Czech, Chinese (Simplified), French, German, Hindi, Indonesian, Italian, Japanese, Korean, Russian, Spanish, Swedish, and Turkish. The summarization prompt instructions were created by translating the original English summarization prompt into all target languages, with all translations checked by proficient speakers of each respective language to ensure linguistic accuracy and cultural appropriateness.

Content filtering and quality control We applied comprehensive filtering criteria to ensure high-quality multilingual content suitable for cross-lingual evaluation:

- Language-based filtering: Using language identification[6], we omitted reviews in languages not covered by the sub-task and retained only products/venues with reviews in more than one languages.
- Content length filtering: Reviews shorter than 50 characters (Amazon) or 20 characters (Google Maps) were removed as non-informative. We applied IQR-based outlier removal per language to eliminate excessively long individual reviews, while enforcing a 1,500-character limit on the final merged multi-document input for manageable human evaluation.
- Language pair balancing: We removed over-represented language combinations to maintain dataset balance and promote multilingual scenarios. We implemented a mixed-content counting algorithm that handles both alphabetic and logographic writing systems appropriately.

Balanced sampling To ensure equal representation of each target language while maximizing data diversity, we implemented a two-stage sampling approach, which first maximizes coverage across unique data items, then achieves exactly 350 examples per target language (4,900 total examples). We prioritized examples without English input to promote true cross-lingual scenarios for less-explored languages.

Data characteristics The final dataset contains 1.1M words across all examples, with an average of 230 words per example. it exhibits strong cross-lingual properties: 86.3% of examples require summarization in a target language different from any of the input languages, and 46.8% contain no English in the source reviews. The dataset comprises 66.0% Google Maps restaurant reviews (3,232 examples) and 34.0% Amazon product reviews (1,668 examples).

2.4 Machine Translation

The MT sub-task adopts the WMT25 General MT test set; full details on data sourcing, difficulty sampling, and human references collection are documented in Kocmi et al. (2025a).

[6] github.com/saffsd/langid.py

Sources and domains Source documents were collected across six domains (news, social, speech, literary, educational, dialogue) and three source languages (Czech, English, Japanese). Speech includes source audio with ASR transcripts, and social includes thread screenshots, with the objective of looking at some of the impacts of multimodal translation. The focus is on the most recent data possible to minimize potential overlap with the pretraining and fine-tuning data of the models under evaluation. All source texts were originally authored in the source language. This approach is crucial to avoid "translationese" in the source texts, which can negatively affect evaluation accuracy (Toral et al., 2018; Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020). To balance domains and source languages, for each domain and source language combination, we targeted ∼9k words and 60–100 segments, with an average segment length of ∼100 words. This design enables the micro-averaging of results across languages and domains without any single category disproportionately influencing the final scores. However, there are some exceptions, as keeping these variables fixed was impractical. For example, the average segment length for the English and Japanese Speech data is 145.27 and 180.59 words, respectively, which is higher than the 100-word objective. Similarly, the dialogue domain's segments have an average length of 178.8 and 147.3 words, respectively. Comprehensive domain-specific collection procedures and final test set statistics are detailed in Kocmi et al. (2025a).

Translation instructions There is no standardized prompt instructions for WMT machine translation evaluation, various are used, from simple 'Translate into {target_lang}:' to more complex instructions adding additional instructions such as 'Your goal is to accurately convey the meaning and nuances of the original {source_lang} text while adhering to {target_lang} grammar, vocabulary, and cultural sensitivities.' (Deutsch et al., 2025).

For our use case, we extend the instruction to cover more details that human translators are asked for. Furthermore, we modify the instructions for each domain. Detailed prompt instructions are in Table 18.

2.5 LLM-as-a-judge for OEG and XLSum

LLM-as-a-judge has recently emerged as an automated solution to open-ended generation evaluation (Zheng et al., 2023b; Verga et al., 2024). It achieves high correlation with human judgment, but its efficacy for languages other than English remains little known (Son et al., 2024). To evaluate the capabilities of models to perform quality assessment of other LLM outputs, we set up the sub-tasks of LLM-as-a-judge for open-ended generation, cross-lingual summarization, and machine translation, where participating systems run evaluation on system outputs from those sub-tasks.

The LLM judges are given the same instructions provided to human annotators, and are assessed by computing their judgments' correlation to human judgment. To evaluate LLM-as-a-judge for the OEG and XLSum sub-tasks, we take all samples that are evaluated with human annotators and use a prompt instruction to judge the system output on a Likert scale of 1–7. For each system output, we run LLM-as-a-judge separately on different evaluation criteria, guided by a rubric each. Specifically:

- OEG: instruction following, naturalness, and coherence
- XLSum: faithfulness, coverage, naturalness, and coherence

The exact prompt instructions are provided in Appendix B. As human evaluation was available for only a subset of languages and systems, LLM-as-a-judge was tested on the same set of data.

2.6 LLM-as-a-judge for MT

Automatic machine translation evaluation is the catalyst of progress in translation technologies, offering a quick, low-cost signal of quality. Early metrics were string-matching against a reference, such as BLEU or ChrF (Papineni et al., 2002; Popović, 2015), which were replaced by trained metrics, such as COMET or MetricX (Rei et al., 2020; Juraska et al., 2023), and finally LLM-as-a-judge (Kocmi and Federmann, 2023). Even though each replacement increased the correlation with human judgment of translation quality, new concerns have emerged regarding language bias, robustness (Moghe et al., 2025; Zouhar et al., 2024a,b), and self-bias for evaluation (Wataoka et al., 2024; Zheng et al., 2023a; Stureborg et al., 2024). This meta-evaluation of automated metrics is usually handled by the WMT Metrics Shared Task (Lavie et al., 2025; Freitag et al., 2024).

In order to test the capabilities of models to perform as LLM-as-a-judge to judge machine translation, we adjust the GEMBA-DA (Kocmi and Fe-

Model and size	Technical report
AyaExpanse 8B	Cohere Labs et al. (2024)
Command R 7B	Cohere et al. (2025)
EuroLLM (9B)	Martins et al. (2025)
Gemma 3 (12B)	DeepMind et al. (2025)
Llama 3.1 (8B)	Grattafiori et al. (2024)
Mistral (7B)	Mistral et al. (2023)
Qwen 2.5 (7B)	Alibaba et al. (2024)
TowerPlus (9B)	Rei et al. (2025)
AyaExpanse 32B	Cohere Labs et al. (2024)
Claude 4 Sonnet	
Command A (111B)	Cohere et al. (2025)
DeepSeek V3 (671B)	DeepSeek et al. (2024)
EuroLLM (22B)	Martins et al. (2025)
Gemini 2.5 Pro	Google et al. (2025)
Gemma 3 (27B)	DeepMind et al. (2025)
GPT 4.1	
Llama 4 Maverick (400B)	
Mistral Medium	
Qwen3 (235B)	Alibaba et al. (2025)
TowerPlus (72B)	Rei et al. (2025)

Table 4: List of all LLMs evaluated in this work. Unshaded models represent "constrained" models, which are smaller and open weights in contrast to "unconstrained" which do not have any limits on being public or size.

Model	LR	MT	OEG	XLSum
Gemini 2.5 Pro	100%	95%	94%	100%
GPT 4.1	85%	90%	100%	94%
DeepSeek V3	90%	80%	88%	65%
Claude 4	95%	78%	81%	88%
Mistral Medium	70%	75%	75%	82%
Llama 4 Maverick	80%	61%	50%	53%
Qwen3 235B	65%	63%	56%	41%
CommandA	75%	56%	62%	59%
Gemma 3 27B	60%	53%	69%	76%
Gemma 3 12B	55%	42%	44%	71%
AyaExpanse 32B	50%	31%	38%	47%
AyaExpanse 8B	30%	20%	31%	29%
Llama 3.1 8B	40%	17%	25%	18%
CommandR7B	20%	14%	19%	-
Qwen2.5 7B	35%	8%	12%	12%
Mistral 7B	5%	5%	6%	6%
TowerPlus 72B	45%	37%	-	35%
TowerPlus 9B	25%	27%	-	24%
EuroLLM 22B	15%	25%	-	-
EuroLLM 9B	10%	19%	-	-

Table 5: Aggregate results across four sub-tasks, converted into percentile ranking (100%=first).

dermann, 2023) prompt with the latest WMT25 human evaluation instruction. The exact prompt instruction is in the Appendix B.

3 Benchmarked Models

For this shared task, we defined two categories for model participation: constrained with several restrictions on model size and licensing; and unconstrained without any limitations. The same way as the General Machine Translation task (Kocmi et al., 2025a). Specifically, the constrained category is restricted to models with fewer than 20B parameters and requires that models be shared as open weights.

Unfortunately, our shared task did not obtain any (valid) participating systems. However, we collected and benchmarked outputs of popular models. The selection process was to identify the strongest-performing system per category for each of the popular model families. This approach ensured that both constrained and unconstrained models were consistently represented; the resulting model list thus reflects a broad yet balanced selection of models, enabling multilingual assessment of the current LLM landscape across languages and problems.

The list of all systems is in Table 4. During the output collection, we ran into budget and API throttling restrictions and thus could not collect some of the systems' outputs for all sub-tasks. When collecting outputs, we set the temperature to 0 and used a unified script.[7]

4 Results

In this section, we present the results and key insights for each sub-task and benchmarked model. Although automatic evaluation was applied to all prompts and outputs, human evaluation was not conducted for all tasks, systems, or languages, due to budget constraints and annotator availability. Nonetheless, human evaluation often proved more reliable than automatic metrics, so we release all annotations for future work on meta-evaluation.

4.1 Linguistic Reasoning

In order to evaluate linguistic reasoning, we choose to break them as much as possible into tasks so that we can grade LLM answers as precisely as possible (which distinguishes this work from previous linguistic reasoning benchmarks). Depending on the task type, we choose either exact match (classification, mapping, fill-in-blanks, editing) or ChrF (translation) as a metric. The scores have to be taken with a grain of salt because ChrF is likely not perfectly expressing the degradations between useless and perfect translation. We assume it rather overestimates translation quality compared to IOL judges. Each task comes with a number of points ([0.5, 1.0, 1.5, 2.0, 2.5]), summing to 20

[7] github.com/wmt-conference/wmt-collect-translations

Model	Average	Spanish	Portuguese	English	French	German	Dutch	Average	Russian	Swedish	Japanese	Ukrainian	Korean	Czech	Estonian	Persian	Chinese
Gemini 2.5 Pro	36.3	40.3	38.1	38.3	39.9	37.3	37.5	36.3	39.5	35.5	35.4	33.9	40.9	36.8	32.5	29.4	28.8
Claude 4	29.7	33.8	35.5	24.6	32.9	33.8	27.9	29.7	30.8	28.4	29.8	32.1	26.3	29.3	27.0	26.3	26.2
DeepSeek V3	23.6	28.5	27.9	23.2	27.9	28.2	24.1	23.6	22.4	21.8	23.1	22.9	20.4	21.5	24.5	20.6	17.4
GPT 4.1	23.4	29.0	27.9	20.5	27.4	24.3	27.1	23.4	24.1	22.7	21.6	23.0	15.3	24.6	29.4	21.8	12.7
Llama 4 Maverick	22.9	30.5	27.2	22.4	26.5	25.9	23.2	22.9	24.6	20.8	20.2	24.5	20.2	22.1	19.4	21.3	14.2
CommandA	19.8	22.0	21.1	17.8	20.6	18.8	23.2	19.8	18.8	17.8	21.3	20.8	21.8	20.1	18.8	18.2	15.7
Mistral Medium	19.8	25.8	23.5	20.4	21.2	24.9	20.8	19.8	21.8	23.2	22.9	16.8	15.2	15.1	14.6	15.7	14.8
Qwen3 235B	17.6	19.9	22.6	22.0	19.1	20.9	21.1	17.6	14.6	21.5	16.0	18.6	17.8	13.0	13.4	14.3	9.6
Gemma 3 27B	17.0	17.1	17.1	18.4	18.3	18.1	17.8	17.0	14.2	19.9	20.0	16.7	18.2	12.5	17.0	14.8	15.0
Gemma 3 12B	16.5	15.6	18.7	21.1	17.3	17.2	18.9	16.5	12.3	15.8	15.9	17.6	12.8	13.3	17.9	16.6	16.0
AyaExpanse 32B	15.3	16.7	17.2	18.9	17.7	14.8	18.7	15.3	19.0	12.3	18.1	10.7	15.7	15.0	3.5	15.1	15.7
TowerPlus 72B	13.4	17.9	17.6	17.6	14.8	11.1	14.4	13.4	14.2	14.6	16.5	13.2	15.1	9.2	13.1	7.9	3.2
Llama 3.1 8B	10.8	14.0	15.5	14.7	16.1	13.1	13.2	10.8	11.1	11.3	10.2	6.0	6.6	6.7	6.4	7.1	10.1
Qwen2.5 7B	10.7	12.6	11.5	13.5	12.2	10.0	10.8	10.7	11.1	7.9	7.7	11.5	11.9	9.6	10.4	8.1	12.2
AyaExpanse 8B	8.7	10.4	13.2	13.5	10.7	10.9	11.8	8.7	7.1	7.4	7.4	7.0	8.4	8.8	1.8	4.4	8.1
TowerPlus 9B	8.5	13.9	6.0	13.8	8.6	13.5	9.8	8.5	7.1	6.8	6.5	3.8	8.8	5.5	9.0	8.0	6.1
CommandR7B	7.3	9.5	8.1	13.5	13.1	11.6	9.2	7.3	8.7	5.9	4.8	5.6	3.2	4.3	0.6	4.6	7.1
EuroLLM 22B	5.7	11.7	7.8	10.4	4.3	8.8	5.9	5.7	6.0	6.7	0.6	6.2	3.7	5.9	4.9	0.0	2.1
EuroLLM 9B	2.6	1.9	3.9	5.7	1.7	1.6	4.9	2.6	0.2	3.9	1.6	2.0	0.7	5.6	4.9	0.0	1.1
Mistral 7B	2.6	6.1	2.7	5.8	2.7	2.2	0.4	2.6	3.4	1.3	2.1	2.1	0.7	1.1	2.4	2.3	3.6

Table 6: Results (number of points) for the linguistic reasoning sub-task (LR) across languages.

points per task and 100 points in total. Points express difficulty, which is not the same across tasks, e.g. translation tasks typically give more points than mapping tasks. The final metric is the sum of prompt-level scores ([0–1]) multiplied by their points, such that the maximum attainable score for each language is 100. The final model ranking is determined by the average number of points across languages. The number of obtained points (out of 100) for each model and language is shown in Table 6. Below are our three key observations.

First, we note that the maximum score in a single language is 40.9 and the maximum average score is 36.3, **indicating headroom** for this kind of task overall. All models failed the majority of tasks. Due to the niche-ness of linguistic reasoning (as opposed to mathematical reasoning), it is unlikely that any of the models has seen very similar tasks during training, which lets this task measure generalization more than memorization. In the 2024 IOL, the winning participant scored 79[8] with human and not automatic scoring, but presented with the same tasks in their mother tongue. The top-scoring model here would have barely made it to a Bronze medal.

Second, the **model ranking is fairly consistent across languages in the top ranks**, with the leading model being Gemini 2.5 Pro across all languages, Claude 4 following in second place, and DeepSeek V3, GPT4.1, and Llama 4 Maverick alternating in place 3. As expected, model size also plays a major role in the ranking: closed-source (presumably large) LLMs are leading in the subtask, followed by CommandA and Qwen3 235B. Notably, Gemma 3 shows good multilingual reasoning performance, with its 27B and 12B versions outperforming TowerPlus at 72B and Aya Expanse at 32B. In the 7–9B range, Llama 3.1 8B is the best. Still, at this model size, we see a steep decline when moving from higher to lower-resource (or unsupported) languages, which is partially due to a lack of instruction following and failing to respond in the required answer format.

Third, most surprisingly, we find that **English is not the language that most models perform strongest in**, although it typically dominates reasoning tasks like math (Chen et al., 2024a). In fact, the "best" solution to the tasks was found by Gemini 2.5 Pro with Korean as the instruction language. In particular, the stronger models show surprising performance drops in English: For Claude, the top performance is 33.8 in German or Spanish, while English lags behind with 24.6 points, scoring the lowest across all languages. Overall, only Gemma 3 12B, Qwen2.5 7B, Aya Expanse 8B, CommandR7B, and EuroLLM 9B performed better in English than all other languages, and in these cases only with a small, perhaps negligible margin.

[8] ioling.org/results/2024

Model	Average	Naturalness	Instruction Following	Coherence
GPT 4.1	6.13	5.94	6.24	6.20
Gemini 2.5 Pro	6.09	5.80	6.25	6.22
DeepSeek V3	5.97	5.65	6.17	6.09
Claude 4	5.96	5.74	6.06	6.08
Mistral Medium	5.96	5.68	6.16	6.03
Gemma 3 27B	5.94	5.59	6.15	6.07
CommandA	5.93	5.65	6.12	6.03
Qwen3 235B	5.90	5.57	6.13	5.99
Llama 4 Maverick	5.89	5.73	6.02	5.93
Gemma 3 12B	5.87	5.57	6.10	5.95
AyaExpanse 32B	5.70	5.33	5.89	5.88
AyaExpanse 8B	5.53	5.10	5.73	5.75
Llama 3.1 8B	5.21	4.82	5.56	5.26
CommandR7B	5.20	4.77	5.47	5.38
Qwen2.5 7B	5.17	4.75	5.40	5.35
Mistral 7B	4.27	3.88	4.49	4.43

Table 7: Per-rubric results for the open-ended generation sub-task. The points are on a Likert-7 scale where 7 is the maximum. See a per-language breakdown in Appendix D Table 19.

Model	Average	Naturalness	Faithfulness	Coherence	Coverage
Gemini 2.5 Pro	6.05	5.90	6.00	6.15	6.13
GPT 4.1	5.99	5.96	5.91	6.17	5.90
Claude 4	5.84	5.76	5.79	5.92	5.86
Mistral Medium	5.77	5.52	5.78	5.82	5.94
Gemma 3 27B	5.73	5.60	5.73	5.85	5.73
Gemma 3 12B	5.68	5.50	5.73	5.84	5.67
DeepSeek V3	5.68	5.35	5.73	5.75	5.88
CommandA	5.64	5.31	5.69	5.77	5.80
Llama 4 Maverick	5.57	5.54	5.54	5.83	5.38
AyaExpanse 32B	5.56	5.46	5.56	5.77	5.44
Qwen3 235B	5.49	5.20	5.45	5.60	5.71
TowerPlus 72B	5.37	5.09	5.42	5.46	5.52
AyaExpanse 8B	5.27	4.81	5.28	5.64	5.34
TowerPlus 9B	4.96	4.82	4.98	5.13	4.92
Llama 3.1 8B	4.49	4.23	4.54	4.62	4.57
Qwen2.5 7B	4.37	3.92	4.48	4.50	4.57
Mistral 7B	3.33	2.77	3.50	3.36	3.70

Table 8: Per-rubric results for the cross-lingual summarization sub-task. The points are on a Likert-7 scale where 7 is the maximum. See a per-language breakdown in Appendix E Table 20.

Explanations for this could be that prompting in other languages brings up the context that is more favorable for solving linguistic reasoning tasks, or that it is just the lack of English dominance in task-relevant data that usually gives it an advantage for other tasks like math or knowledge retrieval. Another explanation could be that model uncertainty might generally be quite high, so that resampling within the same language could cause similar variance as the one we see across languages. We invite future work to dive further into these questions.

4.2 Open-Ended Generation

The open-ended generation sub-task is human-evaluated. We designed a rubric to assess three aspects: instruction following, naturalness, and coherence. This rubric is given to both human evaluators and LLM judges. We only human-evaluated a subset of OEG outputs: 16 systems, 10 languages, and the same 46 questions for all system-language combinations. This is because some questions led to an overly long response, and TowerPlus and EuroLLM models had very high failure rates.

Results are shown in Table 7, with models ranked by their average scores on naturalness, instruction following, and coherence, across all languages. Three points stand out. First, proprietary models generally perform better, except for DeepSeek V3, which is a large open-source mix-of-expert model. Second, performance differences among the leading systems are narrow. Third, naturalness scores show a wider spread than instruction following or coherence, implying a larger gap between the strongest and weakest systems, and highlighting the limitations of systems to produce native sounding text that can be directly used.

4.3 Cross-lingual summarization

We performed a rubric-based human evaluation similar to the setup in OEG. We specifically test for naturalness, faithfulness, coherence, and coverage. We evaluated all 14 target languages in the sub-task. Annotators were proficient in the target language and English but were not expected to speak any other language; therefore, we translated source reviews in all other languages to English using Gemini 2.5 Flash. The user interface allowed them to view the original phrasing of the reviews, if desired.

Three models were excluded from human evaluation for the following reasons: the two EuroLLM models frequently copied input summaries in source languages rather than summarizing them, and CommandR7B had an issue with outputting Polish rather than Czech. Given the novelty and current lack of this type of problem in the field, we conducted the human evaluation for all 17 remaining systems that did not exhibit these evident issues.

Due to budget constraints, we restricted the number of evaluated outputs to 18, resulting in 306 examples rated for each language. The samples were selected based on output diversity using BLEU as the metric. We anticipate that a diverse set of outputs with human ratings will help future efforts in validating automatic metrics for this problem.

We present our preliminary analysis based on 12 target languages in Table 8. When averaged across all target languages, closed-source models have an advantage, with Gemini 2.5 Pro being the best-performing system in the unconstrained track. However, open-weight models are not far behind, led by Gemma3-27B. Model size seems to matter with all of the constrained systems, except for Gemma 3 12B, which punches above its weight, consistently showing lower scores across all languages.

Most models performed well (average rating of 5 and higher) on German, French, Chinese, Italian, Russian, and Spanish. Japanese was the most challenging language. Egyptian Arabic was the most divisive language with 4 clusters, showing a clear advantage of Gemini 2.5 Pro, GPT 4.1, and Claude, with all other models having an average score below 5. Naturalness was the weakest aspect of the generated summaries, often suffering from the models not adhering to the specifically requested language or dialect, or containing untranslated quotes from the source documents.

4.4 Machine Translation

Automatic evaluation We evaluate the MT subtask across 31 language pairs and report AUTORANK, a rank induced by automatic MT metrics where lower is better (1 is best). The AUTORANK is a combination of five different metrics Kocmi et al. (2025a) from three distinct metric families:

- **LLM-as-a-Judge (reference-less).** We use GEMBA-ESA (Kocmi and Federmann, 2023) with two independent judges: GPT 4.1 [9] and CommandA (Cohere et al., 2025), both in a reference-less setting.
- **Trained Reference-based Metrics.** Two supervised metrics trained to approximate human quality judgments with references: MetricX-24-Hybrid-XL[10] (Juraska et al., 2024) and XCOMET-XL[11] (Guerreiro et al., 2024).
- **Trained Quality Estimation (QE).** The reference-less QE metric CometKiwi-XL[12] (Rei et al., 2023), which is also trained to mimic human judgments.

This combination of reference-based and reference-less (or QE) methods is designed to balance their complementary failure modes. Reference-based metrics typically achieve a higher correlation with human judgments when high-quality references are available, while reference-less methods reduce susceptibility to reference bias when references are suboptimal (Freitag et al., 2023). We also account for known issues with specific metrics. To mitigate a common QE pitfall, i.e., being fooled by fluent output in the wrong language, the GEMBA-ESA prompt explicitly specifies the target language.

However, for the two lowest-resource languages in the test set (Bhojpuri and Maasai), we do not apply QE and instead rely solely on `chrF++` (Popović, 2017), computed with `sacrebleu` (Post, 2018). This approach was chosen because the reliability of our main metrics is unestablished for these languages (Falcão et al., 2024; Singh et al., 2024; Wang et al., 2024; Sindhujan et al., 2025), whereas human references required for `chrF++`[13] were available.

The system-level score for each language pair is the average of its paragraph-level (segment-level) scores from each metric across the test set.

Human evaluation The human evaluation is done by Kocmi et al. (2025a) using Error Span Annotation (ESA; Kocmi et al., 2024) and for English to Korean and Japanese to Chinese it relies on Multidimensional Quality Metrics (MQM; Lommel et al., 2014).

The ESA annotators are asked to mark each translation error as well as its severity, "Minor" or "Major". In addition, the annotators are also asked to assign a score from 0 to 100 to the entire annotation segment (usually a paragraph).

In the MQM, annotators are asked to assign categories and subcategories to all error spans. Then, instead of a 0 to 100 slider, the final score is calculated as a sum of error severities, where minor error equals -1 and major error equals -5.

[9] openai.com/index/gpt-4-1
[10] huggingface.co/google/metricx-24-hybrid-xl-v2p6
[11] huggingface.co/Unbabel/XCOMET-XL
[12] huggingface.co/Unbabel/wmt23-cometkiwi-da-xl
[13] SacreBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.5.1`.

Model	Avg. MT AutoRank
Gemini 2.5 Pro	1.02
GPT 4.1	1.51
DeepSeek V3	2.62
Claude 4	2.86
Mistral Medium	3.10
Qwen3 235B	4.10
Llama 4 Maverick	4.34
Gemma 3 27B	4.55
CommandA	4.68
Gemma 3 12B	6.05
TowerPlus 72B	7.00
AyaExpanse 32B	7.32
TowerPlus 9B	8.31
EuroLLM 22B	9.22
AyaExpanse 8B	9.99
EuroLLM 9B	10.60
Llama 3.1 8B	11.81
CommandR7B	11.98
Qwen2.5 7B	14.61
Mistral 7B	18.57

Table 9: Average MT AUTORANK results across language pairs (lower is better). For fairness, all model averages are computed over the same 27 of 31 language pairs, matching Mistral Medium, which lacks outputs for four pairs (see Table 21).

Overall ranking Table 9 reports the average AUTORANK results across the various language pairs. **Gemini 2.5 Pro** leads with an average AUTORANK of **1.02**, followed by **GPT 4.1** (1.51), **DeepSeek V3** (2.62), and **Claude 4** (2.86). This top cluster is clearly separated from a mid-tier (4–6 average ranks; e.g., Qwen3 235B, Llama 4 Maverick, CommandA, Gemma 3 27B) and from compact open-weight models which concentrate above 7–8 on average. **Mistral Medium** remains competitive (3.10), but translating for fewer language pairs than all the other models (27 vs. 31). At the other end, small open-weight baselines (e.g., Qwen2.5-7B, Mistral-7B) cluster around ranks 15–18.

Human evaluation results are in Table 22; due to budget restrictions, not all systems have been evaluated. The overall picture highlights the AUTORANK results. However, we can already see some significant differences showing the limitation of automatic metrics: there is a significant drop in the English to Egyptian Arabic as LLMs mostly output the modern standard Arabic, and DeepSeek significantly underperforms in Serbian, which was not visible on AUTORANK.

Language-pair effects Table 21 in Appendix F reports the fine-grained AUTORANK results across the 31 language pairs. The fine-grained table reveals two consistent trends: (i) High-resource or typologically close directions (e.g., English→German, English→Italian, Japanese→Chinese) yield tight spreads among the strongest systems, often near ranks 1–3. (ii) Low-resource and/or orthography-sensitive directions are much harder. In particular, English→Maasai and English→Bhojpuri show large rank dispersion. Some leaders stay robust (e.g., Gemini 2.5 Pro), while others drop sharply on these pairs (e.g., GPT 4.1 on English→Maasai).

Open vs. closed trends Closed-weight models dominate the top cluster, but **DeepSeek V3** stands out as an open-source mix-of-expert model that competes closely with them. Among mid-sized open models, quality is uneven across language pairs and degrades most on low-resource or script-variant directions.

Relation to other tasks The qualitative picture resembles the pattern in Section 4.1: a tight group of leaders at the top, followed by a broader middle where performance varies more by condition. In MT, the key conditions are the choice of language pairs (especially low-resource and script variants), which ultimately drive the gaps we observe in AUTORANK.

4.5 LLM-as-a-judge for OEG and XLSum

Meta-evaluation of LLM-as-a-judge against humans is a research question in itself. Various correlation techniques are used, e.g., Cohen's Kappa, Kendall Tau, Pearson's, or Spearman's correlations (Liu et al., 2023; Verga et al., 2024). Meta-evaluation in machine translation highlighted many problems of common correlation metrics, such as how handling of ties affects the correlation (Deutsch et al., 2023), how critical grouping of items under Kendall Tau is (Perrella et al., 2024), or why Pearson's correlation may be misleading (Mathur et al., 2020). Thus, we build on top of the MT meta-evaluation research, following the best practices (Freitag et al., 2024).

We anticipate almost no ties in system ranking when all scores are aggregated at the system level, but a large number of ties at the instance level due to our use of rubric scores. Our preliminary data inspection also supports this. We compute two types of correlations at the system and instance levels:

- **Pairwise Accuracy**: pairwise accuracy between

	Pairwise Accuracy	group-by-item acc_{eq} Average	Naturalness	Instruction Following	Coherence
Claude 4	0.95	0.56	0.55	0.57	0.56
GPT 4.1	0.95	0.57	0.54	0.59	0.59
CommandA	0.93	0.57	0.53	0.59	0.59
Qwen3 235B	0.91	0.57	0.53	0.59	0.59
Mistral Medium	0.88	0.55	0.54	0.55	0.56
DeepSeek V3	0.85	0.54	0.50	0.58	0.53
Llama 4 Maverick	0.83	0.53	0.51	0.52	0.56
AyaExpanse 32B	0.73	0.50	0.47	0.53	0.51
Qwen2.5 7B	0.70	0.48	0.46	0.49	0.48
Llama 3.1 8B	0.63	0.44	0.40	0.44	0.47
CommandR7B	0.62	0.49	0.44	0.52	0.51
AyaExpanse 8B	0.58	0.48	0.44	0.51	0.48
Mistral 7B	0.55	0.45	0.40	0.50	0.45

Table 10: System-level (Pairwise Accuracy) and Segment-level (group-by-item acc_{eq} by evaluation criterion) correlation between LLM-as-a-judge and human judgment for OEG.

system ranking and human ranking, neglecting ties.

- **acc_{eq}**: group-by-item pairwise accuracy with ties, then averaged across all items, as introduced by Deutsch et al. (2023). Without losing generality across all sub-tasks, an "item" refers to an input prompt requiring an output in a specific language. We report the results for each evaluation criterion separately, as well as an overall average.

Results for OEG LLM-as-a-judge Table 10 shows both system-level and instance-level accuracy measures between LLM judgment and human judgment. Regarding system ranking pairwise accuracy, the models are roughly split into two groups: LLMs with more than 100B parameters, including both closed-source and open-source ones, achieve high accuracy; small open-source models perform worse, with the lowest performance close to a random toss of a coin of 50%.

Instance-level acc_{eq} scores display a similar overall trend, but top-performing LLM judges are closer to each other. We see that the then top Claude 4 becomes lower than GPT 4.1, CommandA, or Qwen3 235B. The best LLM judge for each criterion also varies: Claude 4 is the best at judging naturalness, but three LLMs, GPT 4.1, CommandA, and Qwen3 235B, achieve the best individual accuracy for judging instruction following and coherence.

	Pairwise Accuracy	group-by-item acc_{eq} Average	Naturalness	Faithfulness	Coverage	Coherence
GPT 4.1	0.93	0.51	0.53	0.50	0.47	0.54
CommandA	0.91	0.50	0.53	0.44	0.50	0.52
Mistral Medium	0.91	0.49	0.51	0.44	0.50	0.52
Llama 4 Maverick	0.89	0.45	0.50	0.41	0.42	0.47
Qwen3 235B	0.89	0.49	0.50	0.47	0.47	0.51
DeepSeek V3	0.87	0.47	0.49	0.46	0.46	0.47
CommandR7B	0.78	0.38	0.35	0.38	0.39	0.40
Qwen2.5 7B	0.78	0.39	0.37	0.36	0.40	0.42
AyaExpanse 32B	0.73	0.40	0.38	0.39	0.42	0.42
Llama 3.1 8B	0.71	0.38	0.38	0.35	0.38	0.42
AyaExpanse 8B	0.69	0.38	0.35	0.38	0.37	0.41
Mistral 7B	0.67	0.36	0.33	0.35	0.39	0.39

Table 11: System-level (Pairwise Accuracy) and Segment-level (group-by-item acc_{eq} by evaluation criterion) correlation between LLM-as-a-judge and human judgment for XLSum.

Results for XLSum LLM-as-a-judge Table 11 shows both system-level and instance-level accuracy measures between LLM judgment and human judgment. At the system level, pairwise accuracy follows a pattern similar to OEG: larger models (CommandA, GPT 4.1, and the 100B+ parameter models) achieve high accuracy between 0.87 and 0.91, while smaller open-source models below 10B parameters perform substantially worse, with accuracies between 0.71 and 0.80.

However, instance-level acc_{eq} scores reveal more concerning patterns. Overall correlations are lower than in OEG, with the best average scores around 0.50. GPT 4.1 demonstrates particularly severe overscoring tendencies, systematically assigning perfect scores to almost all outputs of 4–9 models across all criteria. The best-performing judge also varies considerably by criterion: CommandA achieves the highest accuracy for naturalness and coverage, while GPT 4.1 performs best on faithfulness and coherence despite its overscoring behavior. These patterns suggest that the system-level correlations may reflect spurious text properties rather than the intended evaluation criteria, raising questions about the validity of LLM-as-a-judge for this task.

4.6 LLM-as-a-judge for MT

The meta-evaluation of LLM-as-a-judge was collected in the same way as this year's metric shared task Lavie et al. (2025). Correlations are computed

Model	Avg. SPA	Avg. acc_{eq}
GPT 4.1	0.83	0.49
Claude 4	0.82	0.36
CommandA	0.80	0.39
DeepSeek V3	0.79	0.37
Qwen3 235B	0.78	0.38
AyaExpanse 32B	0.73	0.28
Llama 4 Maverick	0.72	0.19
Qwen2.5 7B	0.67	0.36
Llama 3.1 8B	0.66	0.28
CommandR7B	0.58	0.26
AyaExpanse 8B	0.58	0.22
Mistral 7B	0.54	0.29

Table 12: System-level (Pairwise Accuracy) and Segment-level (acc_{eq}) correlation between LLM-as-a-judge and human judgment for machine translation. Correlations have been averaged across translation directions. Full results are reported in Tables 23 and 24.

at the system level using Pairwise Accuracy (PA, Kocmi et al., 2021) and at the segment level using Pairwise Accuracy with Tie Calibration (acc_{eq}, Deutsch et al., 2023).

We report the average correlations between LLM judges and human annotators in Table 12. At the system level, results resemble those reported in OEG LLM-as-a-judge (Section 4.5): the models are split into two groups, with closed-source and very large models (100+ billion parameters) achieving higher SPA scores ($\geq$ 0.78), while smaller ones range from 0.54 to 0.73. The only outlier is Llama 4 Maverick, which performs poorly compared to similar-sized models, placing in the group of smaller LLMs.

At the segment level, results align with the system-level ones, with models again splitting into the same two performance-based groups. However, two models stand out relative to their peers: GPT 4.1 achieves an acc_{eq} score of 0.49, outperforming all others by a clear margin. Similarly, Qwen2.5 7B reaches 0.36 in terms of acc_{eq}, placing it closer to larger models than to others of comparable size.

Finally, we highlight CommandA, as a dense model with 111B parameters, surpasses several larger MOE competitors such as DeepSeek V3 and Qwen3 235B in Pariwise Ranking and ranks second in acc_{eq}.

5 Conclusion

We introduced the WMT25 Multilingual Instruction Shared Task, where the main contribution is a unified benchmark spanning five evaluation tasks: machine translation, linguistic reasoning, open-ended generation, cross-lingual summarization, and LLM-as-a-judge. The benchmark covers up to 30 languages evaluated both automatically and by humans, and emphasizes robust evaluation of multilingual LLM capabilities. We release all prompts, outputs, and human annotations to facilitate reproducibility and research.

- **Substantial headroom in linguistic reasoning.** Across languages, the best systems achieve well below half of the attainable LR points, indicating that current models struggle with structured, language-agnostic reasoning rather than knowledge recall.
- **English is not always the easiest instruction language.** Several leading models reach their top LR scores in non-English (e.g., Korean, German, Spanish), with noticeable drops in English, suggesting prompting language effects that merit further study.
- **Naturalness is the bottleneck for generation.** In OEG human evaluation, score spread is widest for *naturalness* compared to *instruction following* and *coherence*, echoing user reports that non-English outputs often sound robotic or translationese.
- **Closed-weight models lead, but strong open models follow closely.** Aggregate results and MT AUTORANK ranks show a top cluster of proprietary models, with large open models competitive on several tasks and language pairs.
- **MT quality varies sharply by pair and script.** High-resource or typologically close pairs exhibit tight spreads among top systems, while low-resource and script-variant directions show large gaps and instability.
- **LLM-as-a-judge correlates well at the system level, unevenly at the instance level.** Larger models achieve higher system-level accuracy in OEG/XLSum/MT, while smaller models are not suited for the task.
- **Evaluation reliability still hinges on humans.** Automatic scores enable broad coverage, but human annotations exposed language/script biases, instruction-following failures, and cases where metrics or judges disagree, underscoring the value of our released human-rated subsets.

6 Limitations

Budget-driven coverage limits and occasional model unavailability led to uneven per-task participation. Furthermore, human evaluation was

performed on a subset of the samples.

While we usually report aggregate results across all languages (or language pairs), not all models are trained for all languages. This analysis inevitably penalizes them if some languages are unsupported. Practitioners can refer to the raw data for performance in individual languages of interest.

7 Acknowledgements

We thank many people for their help with annotations, reviewing the linguistic reasoning subtask, and providing or reviewing translations of instructions. They are: Özge Agca, Sweta Agrawal, Ondřej Bojar, Rawan Essam Bondok, Philipp Burlakov, Daryna Dementieva, TG Gowda, Alexandra Grieva, HyoJung Han, Katsuki Isobe, Gunny Kim, Ivan Korovin, Nalin Kumar, Mateusz Lango, Ariston Lim, Jiaming Liu, Andrei Manea, Sourabrata Mukherjee, Youssef Nafea, Ricardo Rei, Steinþór Steingrímsson, Hanka Štěříková, Dušan Variš, Gianluca Vico, and Razan Dyas Wibowo.

We thank Apify and Julian McAuley for their support in obtaining the cross-lingual summarization data.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

Pinzhen Chen is supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546].

Patrícia Schmidtová is supported by Charles University projects GAUK 252986 and SVV 260 698.

References

Sanchit Ahuja, Varun Gumma, and Sunayana Sitaram. 2024. Contamination report for multilingual benchmarks. *arXiv preprint arXiv:2410.16186*.

Team Alibaba, An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Team Alibaba, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Andrew Michael Bean, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kranti Chalamalasetti, Gabriel Bernier-Colborne, Yvan Gauthier, and Sowmya Vajjala. 2025. Test set quality in multilingual LLM evaluation. *arXiv preprint arXiv:2508.02635*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024a. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016. Association for Computational Linguistics.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024b. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726. Association for Computational Linguistics.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, Raphaël Avalos, and others. 2025. Command a: An enterprise-ready large language model. *CoRR*.

Team Cohere Labs, John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

María Andrea Cruz Blandón, Jayasimha Talur, Bruno Charron, Dong Liu, Saab Mansour, and Marcello Federico. 2025. MEMERAG: A multilingual end-to-end meta-evaluation benchmark for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22577–22595. Association for Computational Linguistics.

Team DeepMind, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

DeepSeek, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and others. 2024. DeepSeek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929. Association for Computational Linguistics.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565. ELRA and ICCL.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.

Team Google, Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. Do large language models have an English accent? evaluating and improving the naturalness of multilingual LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings*

of the Eighth Conference on Machine Translation, pages 756–767. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and others. 2025a. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and others. 2025b. Preliminary ranking of WMT25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. 2025. Déjà vu: Multilingual LLM evaluation through the lens of machine translation evaluation. *arXiv preprint arXiv:2504.11829*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Kanojia Diptesh, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172. European Association for Machine Translation.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–Machine parity in language translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, and others. 2025. EuroLLM: Multilingual language models for Europe. *Procedia Computer Science*, 255:53–62.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

Mistral, AQ Jiang, and others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual LLMs. *arXiv preprint arXiv:2506.17080*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459. Association for Computational Linguistics.

Anushka Singh, Ananya Sai, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. How good is zero-shot MT evaluation for low resource Indian languages? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649. Association for Computational Linguistics.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024. MM-Eval: A multilingual meta-evaluation benchmark for LLM-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *Preprint*, arXiv:2405.01724.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning. *Preprint*, arXiv:2409.12126.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516. Association for Computational Linguistics.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

A LLM-in-the-loop PDF Parsing

As described in Section §2.1, the data for 14 of the 15 linguistic exams was extracted using an LLM-in-the-loop pipeline. This approach leveraged the manually parsed English data as a reference to efficiently scale the extraction process, which was followed by human verification and editing. Concretely, we prompted Gemini 2.5 Pro to structure each translated PDF's content into a JSON object by mimicking the provided English example. The LLM was given the unparsed English and translated PDFs, along with the reference JSON object from the English version as a string input.

Prompt Development and Iterations The exact prompt used is shown in Figure 13. As seen, the JSON extraction proved to be highly demanding, requiring the LLM to simultaneously parse content from PDF documents, process long-context inputs, and generate a structured output that must be syntactically valid and programmatically parsable. To improve reliability of the LLM automation process, we iterated on our approach by:

- **Adding step-by-step** instructions to the prompt.
- **Breaking the task down** to parse one problem at a time, which drastically reduced JSON validation failures. This means that for each exam we need at least 5 calls to an LLM (one for each language problem).
- **Implementing a resampling strategy** that only accepted outputs confirmed to be parsable JSON. Figure 14 shows the number of samples drawn for each task/problem language to arrive at a valid JSON. For most problems, a single API call was sufficient. However, the Dâw and Yanyuwa tasks, which employed more complex JSON structures, consistently required more attempts, with some outlier cases, such as for Persian, requiring as many as 50 calls to successfully generate a valid JSON object.

After parsing the PDFs to JSON, we imported the resulting data into spreadsheets so that native speakers could verify its correctness.

```
You are given:
*   An English Linguistic Exam (PDF) with its
    solutions (PDF).
*   The JSON representation of the English
    exam (referred to as "English JSON").
*   The {language} version of the exam (PDF) and its
     solutions
    (PDF).

Objective: Generate a JSON object for the {
    no_problem}
problem ("{problem_language}") of the {language}
    exam.
This JSON should follow
the structural format of the English JSON.

Steps:
1.  Target the "{problem_language}" Problem:
    Isolate the "{problem_language}" problem data.
2.  Structural Template: Use the "{problem_language
    }"
    problem section from the English JSON as the
    structural basis for your new {language} JSON.
3.  Field Handling:
    *  Copy Directly from English JSON: For the
      "{problem_language}" problem, copy the values
          of these
      fields from the English JSON: `Identifier`, `
          Points`,
      `Work Language`, `Task Type`, `Eval Type`, and
          `Task
       Meta`.
    *  Extract from {language} PDFs: For all
       remaining fields,
       populate them with the corresponding content
          extracted
       from the "{problem_language}" section of the
          {language}
       PDFs.
    *  Content Adaptation: The English JSON models
       how PDF
       content should appear in the JSON. If its
          content isn't
       a direct PDF copy (e.g., it's formatted/
          structured),
       then similarly adapt the {language} PDF
          content to match
       this presentation style and any processing
       evident in the English JSON.

Attached are the PDFs, here is the English JSON:
```
{json_object}
```
Make sure all values in the JSON have the same
    length and
that the JSON itself is parsable with json.loads()
    in Python.
Output only the JSON object and nothing else.
```

Table 13: Prompt used to extract a JSON object from a translated PDF by mimicking the structure of a manually parsed example ("json_object") from the original English PDF.

Language	Koryak	Hadza	Komnzo	Dâw	Yanyuwa
Ukrainian	1	1	1	2	4
Russian	1	1	1	4	18
Portuguese	1	1	3	11	11
Czech	1	1	1	1	11
Spanish	1	1	1	2	1
Swedish	1	1	1	1	1
Chinese	1	1	1	1	2
Korean	1	1	1	1	5
Persian	1	1	1	1	52
French	1	1	1	1	2
German	1	1	1	16	14
Japanese	1	1	1	1	4
Estonian	1	1	1	1	1
Dutch	1	1	3	3	9

Table 14: Number of API calls required to extract a parsable JSON object varied by task and language.

B LLM-as-a-judge Instructions

LLM-as-a-judge prompts for OEG, MT, and XLSum are in Tables 15 to 17 respectively.

```
Score the response generated by a system to a user's request in {language} on a likert scale from 1 to 7. The quality levels associated with numerical scores are provided below:
7: The response fully adheres to all instructions that the user provided.
5: The chatbot mostly followed the instructions, conforming to the main points of the request but missing some details.
3: The chatbot followed only a small portion of the instructions or missed important points.
1: The chatbot entirely disregarded the instructions; the response appears to have no relation to the user's request or is not in {language}.
Return only the score and nothing else.
Prompt: {input}
Response: {response}
Score (1-7):
```

```
Score the response generated by a system to a user's request in {language} on a likert scale from 1 to 7. The quality levels associated with numerical scores are provided below:
7: The response represents fluent {language} text that might have been written by a native human speaker.
5: The response has some disfluencies that are noticeable but don't significantly hinder understanding the response.
3: The response is highly disfluent. There are several grammatical errors. Most of the meaning can be determined, but only with conscious effort.
1: The response is incomprehensible or is not in {language}.
Return only the score and nothing else.
Prompt: {input}
Response: {response}
Score (1-7):
```

```
Score the response generated by a system to a user's request in {language} on a likert scale from 1 to 7. The quality levels associated with numerical scores are provided below:
7: The response is logically sound and appropriately structured with a clear sequence of nicely connected ideas and topics with no leaps in reasoning.
5: The response is generally well-structured and has a generally clear overall progression of ideas, but introduces a few logical gaps, or suddenly switches topics without an appropriate transition.
3: The response lacks an overall flow, and/or has multiple noticeable jumps between topics. It is possible to discern some relevant ideas, but the overall purpose of the response is incoherent.
1: The response has no overall structure, is in no way logically sound, and/or can be divided into many mostly-unrelated sections. It is difficult to identify any points the text is trying to make.
Return only the score and nothing else.
Prompt: {input}
Response: {response}
Score (1-7):
```

Table 15: Prompt instructions used in the LLM-as-a-judge for OEG sub-task.

```
Score the following translation from {source_lang} to {target_lang} on a scale from 0 to 100, where a score of 0 means a broken or poor translation; 33 indicates a flawed translation with significant issues; 66 indicates a good translation with only minor issues in grammar, fluency, or consistency; and 100 represents a perfect translation in both meaning and grammar. Answer with only a whole number representing the score, and nothing else.
{source_lang} source text:
{source_seg}
{target_lang} translation:
{target_seg}
```

Table 16: Prompt instructions used in the LLM-as-a-judge for MT sub-task.

```
Score the summary generated by a system based on a set of reviews in {language} on a likert scale from 1 to 7. Evaluate whether all information in the summary can be traced back to the reviews. Treat the reviews as the source of truth and do not consider any external information. The quality levels associated with numerical scores are provided below:
7: All of the information in the summary is fully supported by the reviews and no meaning was changed.
5: Most information is supported, but a small part of the summary contains information that either contradicts or cannot be verified by the reviews.
3: More than half of the information in the summary either contradicts or cannot be verified by the reviews.
1: The summary is fully made up of information that either contradicts or cannot be verified by the reviews.
Return only the score and nothing else.
Reviews: {input}
Summary: {response}
Score (1-7):
```

```
Score the summary generated by a system based on a set of reviews in {language} on a likert scale from 1 to 7. Read the reviews and identify the most important points, then evaluate whether these key points are covered by the summary. The quality levels associated with numerical scores are provided below:
7: The summary covers all key points.
5: The summary covers about two thirds of the key points.
3: The summary covers about a third of the key points.
1: The summary does not cover any of the key points mentioned in the reviews.
Return only the score and nothing else.
Reviews: {input}
Summary: {response}
Score (1-7):
```

```
Score the summary generated by a system based on a set of reviews in {language} on a likert scale from 1 to 7. Evaluate the degree to which the summary appears to be fluent, natural text in {language}, that is appropriate in terms of tone and formality. The quality levels associated with numerical scores are provided below:
7: The summary represents fluent {language} text that might have been written by a native human speaker.
5: The summary has some disfluencies that are noticeable but don't significantly hinder understanding the summary.
3: The summary is highly disfluent. There are several grammatical errors. Most of the meaning can be determined, but only with conscious effort. Alternatively, there are some words in a foreign language.
1: The summary is incomprehensible, or is not in {language}.
Return only the score and nothing else.
Reviews: {input}
Summary: {response}
Score (1-7):
```

```
Score the summary generated by a system based on a set of reviews in {language} on a likert scale from 1 to 7. Evaluate the degree to which the summary appears to be logically sound and internally consistent. The quality levels associated with numerical scores are provided below:
7: The summary is logically sound and appropriately structured with a clear sequence of nicely connected ideas and topics with no leaps in reasoning.
5: The summary is generally well-structured and has a generally clear overall progression of ideas, but introduces a few logical gaps, or suddenly switches topics without an appropriate transition.
3: The summary lacks an overall flow, and/or has multiple noticeable jumps between topics. It is possible to discern some relevant ideas, but the overall purpose of the summary is incoherent.
1: The summary has no overall structure, is in no way logically sound, and/or can be divided into many mostly-unrelated sections. It is difficult to identify any points the text is trying to make.
Return only the score and nothing else.
Reviews: {input}
Summary: {response}
Score (1-7):
```

Table 17: Prompt instructions used in the LLM-as-a-judge for XLSum sub-task.

C MT Prompt instructions

```
You are a professional {source_language}-to-{
    target_language} translator, tasked with
    providing translations suitable for use in {
    target_region} ({tgt_language_code}). Your goal
     is to accurately convey the meaning and
    nuances of the original {source_language} text
    while adhering to {target_language} grammar,
    vocabulary, and cultural sensitivities. The
    original {source_language} text {
    domain_description}. {domain_instruction}
    Produce only the {target_language} translation,
     without any additional explanations or
    commentary. Retain the paragraph breaks (double
     new lines) from the input text. Please
    translate the following {source_language} text
    into {target_language} ({tgt_language_code}):\n
    \n{input_text}
```

```
news: Ensure the translation is formal, objective,
    and clear. Maintain a neutral and informative
    tone consistent with journalistic standards.
social: Ensure you do not reproduce spelling
    mistakes, abbreviations or marks of
    expressivity. Platform-specific elements such
    as hashtags or userids should be translated as-
    is.
literary: Aim to maintain the original tone and
    register, retaining the emotional depth of the
    story. Dialogues should sound natural and
    follow the conventions of the target language.
speech: Pay attention to errors that mimic speech
    transcription errors and fix as necessary.
    Maintain the flow and colloquial style of the
    speaker in the translation.
edu: Preserve the line breaks. Use precise
    terminology and a tone appropriate for academic
     or instructional materials.
dialogue: Maintain dialog turn structure and speaker
     indicators (X, Y). Ensure natural flow,
    consistent tone (feminine/masculine, polite/
    familiar), and preserve any HTML tags (e.g.,
    italics).
```

Table 18: Prompt instruction used in the machine translation sub-task together with domain information.

D Open-Ended Generation Results by Language

Table 19 details the open-ended generation performance in each language (and locale).

E Cross-lingual Summarization Results by Language

Table 20 details the cross-lingual summarization performance in each language.

F Machine Translation Fine-Grained Results

Table 21 reports the fine-grained MT AUTORANK scores for all models by language pair.

Model	*Egyptian Arabic*	*Bengali*	*Simplified Chinese*	*Czech*	*English*	*German*	*Hindi*	*Indonesian*	*Japanese*	*Russian*
GPT 4.1	5.81	4.22	6.21	6.38	6.49	6.64	6.16	6.66	6.00	6.69
Gemini 2.5 Pro	5.45	4.67	6.47	6.32	6.24	6.34	6.31	6.21	6.22	6.64
DeepSeek V3	5.12	4.44	6.36	6.14	6.36	6.40	6.26	5.96	6.09	6.54
Claude 4	5.84	4.23	6.11	5.70	6.37	6.54	6.04	6.26	6.01	6.50
Mistral Medium	4.98	4.32	6.25	6.05	6.28	6.54	6.14	6.38	6.05	6.59
Gemma 3 27B	5.32	4.33	6.19	5.93	6.23	6.41	6.13	6.20	6.02	6.61
CommandA	5.92	3.83	6.24	5.96	6.43	6.46	5.78	6.38	5.90	6.43
Qwen3 235B	5.04	4.17	6.42	5.84	6.43	6.25	6.11	6.18	5.91	6.62
Llama 4 Maverick	5.57	4.23	5.96	5.85	6.42	6.36	6.04	6.43	5.59	6.49
Gemma 3 12B	5.41	4.15	6.12	5.84	6.18	6.36	5.99	6.15	5.86	6.67
AyaExpanse 32B	4.67	3.35	5.95	5.92	6.39	6.51	5.98	6.18	5.72	6.32
AyaExpanse 8B	4.47	3.06	5.89	5.65	6.29	6.17	6.08	5.86	5.59	6.20
Llama 3.1 8B	3.81	3.12	5.62	5.04	6.33	5.98	5.80	6.00	4.59	5.86
CommandR7B	4.32	3.04	5.87	4.75	6.30	6.19	5.64	5.33	5.50	5.09
Qwen2.5 7B	4.13	3.15	5.95	4.35	6.36	6.14	4.86	5.45	5.53	5.75
Mistral 7B	1.79	1.80	5.29	4.27	6.36	5.86	3.57	4.65	3.83	5.25

Table 19: Human evaluation results for open-ended generation by language-locale. The scores are averaged across all evaluated rubrics.

Model	Czech	Egyptian Arabic	French	German	Hindi	Indonesian	Italian	Japanese	Korean	Russian	Simplified Chinese	Spanish	Swedish	Turkish
Gemini 2.5 Pro	6.11	5.57	6.42	6.51	6.22	5.81	6.11	5.44	5.74	6.42	6.49	5.79	6.07	6.00
GPT 4.1	6.00	5.56	6.38	6.31	5.93	5.79	6.06	5.71	5.58	6.54	6.33	5.68	6.08	5.88
Claude 4	5.60	5.31	6.33	6.54	5.88	5.44	6.06	5.96	5.51	6.32	6.47	5.86	4.62	5.79
Mistral Medium	5.58	4.94	6.06	6.36	5.93	5.56	5.88	5.32	5.49	6.18	6.38	5.71	5.50	5.88
Gemma 3 27B	5.35	4.82	6.42	6.28	5.76	5.01	5.99	5.64	5.29	6.18	6.22	5.69	5.71	5.82
Gemma 3 12B	5.36	4.69	5.88	6.32	5.83	5.64	6.17	5.56	5.32	6.21	6.22	5.50	5.14	5.72
DeepSeek V3	5.47	4.90	6.00	6.25	6.07	5.40	5.92	5.24	5.46	6.03	6.38	5.57	5.00	5.81
CommandA	5.07	4.82	6.31	6.15	5.65	5.36	5.89	5.53	5.57	5.92	6.08	5.60	5.18	5.88
Llama 4 Maverick	5.39	4.71	6.25	5.86	5.90	5.28	5.46	5.26	5.57	5.67	5.99	5.54	5.42	5.71
AyaExpanse 32B	5.31	4.89	5.86	6.44	5.60	5.88	5.53	5.51	5.12	6.08	5.78	5.61	4.36	5.82
Qwen3 235B	4.46	4.12	5.92	6.39	5.88	5.42	5.94	5.15	5.12	6.17	6.47	5.72	4.53	5.54
TowerPlus 72B	5.36	2.49	6.31	5.60	5.69	5.68	5.79	5.22	5.21	5.81	6.18	5.58	5.17	5.12
AyaExpanse 8B	5.31	4.49	6.22	5.68	5.72	5.11	5.97	4.93	4.97	5.24	5.51	5.28	3.67	5.65
TowerPlus 9B	5.29	2.28	5.74	5.54	5.64	4.12	5.78	4.82	5.00	5.26	5.94	5.62	4.96	3.49
Llama 3.1 8B	3.57	2.71	5.68	4.92	4.76	4.51	5.21	3.58	3.97	5.14	5.15	5.04	4.07	4.54
Qwen2.5 7B	3.21	2.65	5.76	5.21	3.65	4.69	4.58	4.40	4.50	4.65	5.26	5.43	3.04	4.06
Mistral 7B	2.35	1.57	4.76	3.43	2.79	3.69	4.88	2.68	3.18	4.65	3.83	3.89	2.07	2.90

Table 20: Human evaluation results for cross-lingual summarization by language. The scores are averaged across all evaluated rubrics.

Model	Czech→German	Czech→Ukrainian	English→Arabic	English→Bhojpuri	English→Bengali	English→Czech	English→German	English→Greek	English→Estonian	English→Persian	English→Hindi	English→Indonesian	English→Icelandic	English→Italian	English→Japanese	English→Kannada	English→Korean	English→Lithuanian	English→Masai	English→Marathi	English→Romanian	English→Russian	English→Serbian Cyr.	English→Serbian Lat.	English→Swedish	English→Thai	English→Turkish	English→Ukrainian	English→Vietnamese	English→Chinese	Japanese→Chinese
Gemini 2.5 Pro	1.1	1.0	1.0	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.2	1.0	1.0	1.0	1.0	6.2	1.0	1.0	1.0	1.0	1.1	1.0	1.0	1.0	1.0	1.0	1.0	1.0
GPT 4.1	1.0	1.1	1.6	3.5	1.7	1.3	1.0	1.3	1.3	1.4	1.5	1.5	1.1	1.0	1.1	3.8	1.2	1.4	19.0	2.4	1.1	1.4	1.2	1.1	1.5	1.7	1.2	1.0	1.1	1.4	1.6
DeepSeek V3	1.7	1.9	1.2	3.3	2.5	1.8	1.1	3.6	5.1	2.1	1.7	1.9	6.6	2.0	2.0	4.4	2.4	5.7	6.3	3.0	2.0	1.6	4.5	1.0	2.0	2.0	1.4	1.9	1.4	3.2	3.2
Claude 4	2.3	2.1	2.3	3.0	2.4	3.7	2.4	2.5	3.1	2.8	3.0	3.3	3.4	3.7	2.3	3.2	2.0	3.2	1.0	3.0	3.4	3.2	3.4	2.9	2.8	2.7	2.9	3.1	2.6	3.0	2.8
Mistral Medium	1.9	2.3	1.4	5.5	2.2	2.7	1.2	2.8	7.2	2.4	4.3	2.5	6.5	2.5	2.2	4.3	2.9	6.4	–	4.0	2.6	–	–	–	2.3	2.4	2.6	2.5	1.7	1.5	2.8
Qwen3 235B	4.8	5.2	3.3	6.7	3.9	4.9	2.8	4.3	7.3	5.1	4.3	1.9	8.9	2.5	3.1	4.5	3.1	4.5	1.7	5.0	3.1	3.2	6.6	4.3	4.6	1.9	4.2	4.4	1.4	1.4	3.3
Llama 4 Maverick	3.8	3.8	4.1	4.1	3.6	4.9	3.2	3.8	4.0	3.3	4.3	4.1	5.8	9.6	3.9	5.6	4.5	4.0	2.0	4.7	3.5	5.1	5.4	4.2	3.4	3.0	4.0	4.1	3.7	3.5	7.0
CommandA	2.2	2.5	2.8	4.1	6.5	3.6	1.8	2.4	8.4	2.9	3.6	4.3	10.7	3.4	3.0	10.1	2.8	6.6	8.8	8.4	2.7	4.4	10.1	5.5	5.1	7.4	4.5	3.3	5.1	4.6	3.6
Gemma 3 27B	3.5	2.7	3.0	5.1	4.8	3.8	10.4	7.0	3.7	2.4	3.2	2.8	6.3	6.1	3.3	5.2	6.8	3.9	16.6	3.2	2.8	3.2	6.7	7.0	2.6	2.4	5.1	5.8	5.9	4.3	6.8
Gemma 3 12B	5.8	4.9	4.5	7.3	4.9	6.0	6.6	5.7	6.3	3.7	4.4	3.6	9.2	8.0	6.2	9.1	4.6	6.1	10.2	7.8	4.4	6.2	6.7	5.6	6.7	5.9	4.9	7.2	4.7	5.4	9.7
TowerPlus 72B	4.6	4.4	6.3	7.6	9.7	5.7	3.9	13.9	10.7	11.2	6.0	5.3	3.7	3.8	3.6	17.1	4.3	14.7	10.1	11.3	5.1	4.1	15.4	10.3	4.2	4.3	8.1	4.9	5.0	4.8	4.8
AyaExpanse 32B	4.0	3.8	1.9	6.8	12.8	4.3	2.7	3.1	17.1	3.7	5.3	4.4	17.5	4.3	4.2	16.3	4.3	13.9	7.7	11.8	3.4	5.6	19.0	11.5	12.9	14.7	5.7	4.2	3.0	6.1	5.4
TowerPlus 9B	5.0	4.0	16.4	7.6	13.3	4.8	3.7	16.4	14.4	13.6	4.8	12.0	2.4	5.0	4.5	11.9	5.0	16.4	5.1	4.8	3.0	4.7	15.7	15.1	3.2	12.0	13.9	3.9	10.6	6.1	5.7
EuroLLM 22B	5.5	4.4	3.1	8.1	20.0	5.5	4.0	3.6	4.0	19.0	6.9	16.9	18.9	4.8	7.5	19.0	6.9	4.8	9.3	11.5	4.4	6.1	12.0	6.8	4.7	19.3	5.0	6.2	20.0	6.8	8.0
AyaExpanse 8B	8.0	6.5	2.4	9.8	17.1	7.2	6.3	4.5	20.0	5.8	7.0	5.4	20.0	7.3	7.0	18.8	6.6	18.3	9.3	13.4	5.3	7.6	17.5	18.4	20.0	17.7	7.8	6.5	4.4	8.3	8.5
EuroLLM 9B	11.5	7.2	5.1	8.7	18.2	8.5	6.3	4.9	6.8	20.0	7.2	20.0	15.6	7.5	10.9	18.8	10.2	5.3	9.2	8.7	5.6	9.6	13.0	8.0	5.6	20.0	6.0	8.6	17.2	9.3	12.5
Llama 3.1 8B	13.6	12.1	10.8	20.0	10.2	13.8	12.3	11.7	13.0	9.2	9.1	8.4	15.8	12.8	11.5	13.6	13.5	14.9	9.0	11.2	10.7	14.1	11.3	11.2	8.7	8.6	11.1	13.0	7.2	10.6	11.4
CommandR7B	9.4	12.3	3.4	9.8	15.3	13.2	8.6	9.0	18.8	8.4	10.5	10.4	18.3	9.5	8.9	16.9	9.8	16.5	3.6	14.0	9.9	19.0	18.6	19.0	17.2	17.4	10.9	15.6	8.0	10.7	10.6
Qwen2.5 7B	17.4	20.0	11.6	10.4	14.9	18.2	13.6	18.7	17.8	14.6	16.5	8.5	19.4	13.2	11.0	20.0	13.6	18.6	9.8	18.4	20.0	11.4	19.0	18.2	16.1	7.9	14.5	20.0	6.5	5.9	7.3
Mistral 7B	20.0	15.9	20.0	12.1	19.1	20.0	20.0	20.0	19.7	19.3	20.0	17.1	19.4	20.0	20.0	19.6	20.0	20.0	13.7	20.0	17.5	18.9	15.9	15.5	12.9	16.6	20.0	15.9	16.2	20.0	20.0

Table 21: Machine translation AUTORANK results across language pairs (lower is better).

Model	Czech→German		Czech→Ukrainian		English→Arabic		English→Bhojpuri		English→Chinese		English→Czech		English→Estonian		English→Icelandic		English→Italian		English→Japanese		English→Korean		English→Maasai		English→Russian		English→Serbian Cyrillic		English→Ukrainian		Japanese→Chinese	
Gemini 2.5 Pro	91	1	93	1	61	1	95	1	84	1	89	1	79	1	78	1	79	1	86	1	-3	1	10	6	83	1	94	1	90	1	-4	1
GPT 4.1	89	1	92	1	77	2	83	4	84	1	81	1	72	1	68	1	79	1	84	1	-3	1		19	76	1	92	1	88	1	-6	2
Claude 4	89	2	89	2	56	2	83	3	87	3	80	4	53	3	48	3	72	4	79	2	-3	2	8	1	76	3	90	3	86	3	-6	3
DeepSeek V3	88	2	89	2	57	1	77	3	85	3	85	2		5		7	72	2	79	2	-4	2	3	6	74	2	79	4	86	2	-8	3
Mistral Medium	87	2	89	2	36	1		6	80	2	80	3		7		6	74	2	85	2	-5	3							85	2	-10	3
CommandA	87	2	86	2	74	3	73	4		5	76	4		8		11	73	3		3	-5	3	1	9		4		10	84	3		4
TowerPlus 9B	80	5	85	4		16		8		6	66	5		14	57	2	61	5		4	-7	5	1	5		5		16	84	4	-13	6
Gemma 3 27B	82	4	89	3		3	56	5		4	76	4	46	4		6		6		3		7		17	62	3		7		6		7
Llama 4 Maverick		4		4		4	76	4	81	4		5		4		6		10		4		4	5	2		5		5	86	4		7
Qwen3 235B		5		5		3		7	84	1		5		7		9	67	2		3		3	3	2	68	3		7		4		3
Gemma 3 12B	77	6		5		4		7		5		6		6	16	9	54	8		6	-6	5	3	10		6	75	7		7		10
EuroLLM 9B		12		7		5		9		9		8		7		16	57	8		11		10	1	9		10	42	13		9		12
Llama 3.1 8B		14		12		11		20		11		14		13	11	16		13		12		14	3	9		14	58	11		13		11
AyaExpanse 8B		8		6	2	2		10		8		7		20		20	57	7		7		7	6	9		8		18		6		8
EuroLLM 22B		6		4		3		8		7		6	47	4		19		5		8		7	0	9		6		12		6		8
TowerPlus 72B		5		4		6		8		5		6		11	46	4		4		4		4		10		4		15		5		5
CommandR7B		9		12	4	3		10		11		13		19		18		10		9		10	2	4		19		19		16		11
AyaExpanse 32B		4		4		2		7		6		4		17		18		4		4		4	3	8		6		19		4		5
Qwen2.5 7B		17		20		12		10		6		18		18		19		13		11		14	3	10		11		19		20		7
Mistral 7B		20		16		20		12		20		20		20		19		20		20		20		14		19		16		16		20

Table 22: Combination of human evaluation of LLMs for machine translation and AUTORANK (Kocmi et al., 2025a,b). The left column shows human evaluation (higher is better) either with ESA or MQM annotation protocol (Kocmi et al., 2024; Freitag et al., 2021) and the right column for each language shows the AUTORANK (lower is better) based on Table 21.

	Czech → German	Czech → Ukrainian	English → Arabic	English → Bhojpuri	English → Czech	English → Estonian	English → Icelandic	English → Italian	English → Japanese	English → Korean	English → Maasai	English → Russian	English → Serbian	English → Ukrainian	English → Chinese	Japanese → Chinese
GPT 4.1	0.90	0.85	0.87	0.78	0.87	0.84	0.93	0.84	0.76	0.88	0.61	0.82	0.89	0.78	0.80	0.94
Claude 4	0.88	0.84	0.73	0.77	0.88	0.79	0.91	0.84	0.75	0.92	0.62	0.77	0.88	0.78	0.85	0.93
CommandA	0.86	0.83	0.74	0.69	0.87	0.78	0.81	0.83	0.70	0.89	0.59	0.81	0.83	0.80	0.83	0.91
DeepSeek V3	0.86	0.85	0.55	0.64	0.85	0.85	0.86	0.83	0.67	0.90	0.61	0.76	0.85	0.81	0.84	0.88
Qwen3 235B	0.84	0.84	0.53	0.61	0.88	0.81	0.84	0.83	0.70	0.89	0.59	0.80	0.84	0.81	0.86	0.81
AyaExpanse 32B	0.79	0.83	0.54	0.64	0.83	0.63	0.65	0.82	0.63	0.90	0.58	0.74	0.69	0.77	0.82	0.79
Llama 4 Maverick	0.82	0.82	0.59	0.66	0.66	0.75	0.78	0.74	0.65	0.82	0.60	0.65	0.74	0.65	0.80	0.84
Qwen2.5 7B	0.72	0.66	0.34	0.54	0.78	0.54	0.66	0.82	0.59	0.80	0.57	0.77	0.74	0.64	0.79	0.78
Llama 3.1 8B	0.75	0.70	0.20	0.58	0.74	0.63	0.74	0.78	0.59	0.80	0.55	0.71	0.77	0.62	0.81	0.63
CommandR7B	0.77	0.71	0.17	0.55	0.68	0.47	0.44	0.80	0.61	0.76	0.57	0.57	0.37	0.50	0.68	0.67
AyaExpanse 8B	0.76	0.66	0.34	0.49	0.70	0.46	0.44	0.72	0.56	0.73	0.50	0.63	0.42	0.51	0.67	0.70
Mistral 7B	0.64	0.56	0.21	0.48	0.62	0.49	0.46	0.63	0.44	0.60	0.56	0.53	0.61	0.55	0.62	0.57

Table 23: System-level Soft Pairwise Accuracy (SPA) computed between the LLM judges and human annotators in the task of machine translation evaluation.

	Czech → German	Czech → Ukrainian	English → Arabic	English → Bhojpuri	English → Czech	English → Estonian	English → Icelandic	English → Italian	English → Japanese	English → Korean	English → Maasai	English → Russian	English → Serbian	English → Ukrainian	English → Chinese	Japanese → Chinese
GPT 4.1	0.46	0.40	0.53	0.56	0.47	0.52	0.66	0.45	0.45	0.50	0.54	0.46	0.51	0.43	0.41	0.46
CommandA	0.41	0.35	0.44	0.34	0.38	0.35	0.43	0.40	0.33	0.47	0.49	0.36	0.43	0.34	0.34	0.38
Qwen3 235B	0.39	0.33	0.37	0.29	0.35	0.33	0.42	0.38	0.34	0.47	0.49	0.36	0.41	0.33	0.36	0.40
DeepSeek V3	0.36	0.33	0.37	0.38	0.31	0.33	0.47	0.34	0.28	0.48	0.49	0.33	0.45	0.31	0.35	0.39
Claude 4	0.35	0.29	0.42	0.35	0.28	0.30	0.52	0.28	0.30	0.48	0.56	0.28	0.42	0.25	0.30	0.39
Qwen2.5 7B	0.36	0.32	0.37	0.31	0.31	0.29	0.38	0.37	0.28	0.47	0.49	0.36	0.37	0.34	0.33	0.36
Mistral 7B	0.28	0.28	0.37	0.22	0.23	0.22	0.24	0.26	0.25	0.46	0.49	0.25	0.27	0.23	0.23	0.30
AyaExpanse 32B	0.28	0.29	0.37	0.27	0.22	0.15	0.21	0.28	0.23	0.46	0.49	0.23	0.27	0.22	0.24	0.33
Llama 3.1 8B	0.27	0.25	0.37	0.20	0.22	0.25	0.22	0.24	0.23	0.46	0.49	0.25	0.26	0.20	0.28	0.31
CommandR7B	0.22	0.21	0.37	0.18	0.19	0.20	0.19	0.26	0.25	0.47	0.49	0.24	0.17	0.22	0.26	0.32
AyaExpanse 8B	0.19	0.20	0.37	0.20	0.12	0.11	0.11	0.18	0.13	0.46	0.49	0.13	0.18	0.12	0.18	0.29
Llama 4 Maverick	0.14	0.17	0.37	0.19	0.05	0.06	0.24	0.11	0.10	0.46	0.49	0.07	0.18	0.06	0.12	0.26

Table 24: Segment-level Pairwise Accuracy with Tie Calibration (acc_{eq} computed between the LLM judges and human annotators in the task of machine translation evaluation.

Findings of the WMT25 Shared Task on Automated Translation Evaluation Systems: Linguistic Diversity is Challenging and References Still Help

Alon Lavie[(1)], **Greg Hanneman**[(2)], **Sweta Agrawal**[(3)], **Diptesh Kanojia**[(4)]
Chi-kiu Lo 羅致翹[(5)], **Vilém Zouhar**[(6)], **Frédéric Blain**[(7)], **Chrysoula Zerva**[(8,9)]
Eleftherios Avramidis[(10)], **Sourabh Deoghare**[(11)], **Archchana Sindhujan**[(4)]
Jiayi Wang[(12)], **David I. Adelani**[(13,14)], **Brian Thompson**[(2)], **Tom Kocmi**[(15)]
Markus Freitag[(3)], **Daniel Deutsch**[(3)]
[(1)]Carnegie Mellon University [(2)]Unaffiliated [(3)]Google [(4)]University of Surrey
[(5)]National Research Council Canada [(6)]ETH Zurich [(7)]Tilburg University
[(8)]Instituto Superior Técnico, Universidade de Lisboa [(9)]Instituto de Telecomunicações
[(10)]German Research Center for Artificial Intelligence (DFKI)
[(11)]Indian Institute of Technology, Bombay [(12)]University College London
[(13)]McGill University [(14)]Mila - Quebec AI Institute [(15)]Cohere
wmt-qe-metrics-organizers@googlegroups.com

Abstract

The WMT25 Shared Task on Automated Translation Evaluation Systems evaluates metrics and quality estimation systems that assess the quality of language translation systems. This task unifies and consolidates the separate WMT shared tasks on Machine Translation Evaluation Metrics and Quality Estimation from previous years. Our primary goal is to encourage the development and assessment of new state-of-the-art translation quality evaluation systems. The shared task this year consisted of three subtasks: (1) segment-level quality score prediction, (2) span-level translation error annotation, and (3) quality-informed segment-level error correction. The evaluation data for the shared task were provided by the General MT shared task and were complemented by "challenge sets" from both the organizers and participants. Task 1 results indicate the strong performance of large LLMs at the system level, while reference-based baseline metrics outperform LLMs at the segment level. Task 2 results indicate that accurate error detection and balancing precision and recall are persistent challenges. Task 3 results show that minimal editing is challenging even when informed by quality indicators. Robustness across the broad diversity of languages remains a major challenge across all three subtasks.

1 Introduction

The WMT25 Shared Task on Automated Translation Quality Evaluation Systems[1] evaluates automated systems for assessing and improving translation quality, including automated quality metrics, quality estimation systems, and quality-informed translation error correction. This task builds on previous years' shared tasks (Freitag et al., 2024; Zerva et al., 2024) and unifies previously separate WMT shared tasks on Machine Translation Evaluation Metrics and Quality Estimation. Automated translation quality evaluation systems play a critical role in the research, development and deployment of machine translation systems, and more recently, of multilingual LLMs. They are also critical components in automated translation workflows for large-scale commercial translation use-cases.

Source	Non parliamo italiano.
Translation	I don't speak Spanish!
Task 1: score prediction	→ 25%
Task 2: error span prediction	→ I don't speak Spanish!
Task 3: post-editing	→ We don't speak Italian.

Table 1: Illustration of the three primary subtasks: segment-level quality score prediction, span-level error detection, and quality-informed post-editing. (The challenge sets subtask is not shown.)

The shared task consists of three primary subtasks shown in Table 1: (1) segment-level quality score prediction, (2) span-level translation error detection, and (3) quality-informed segment-level error correction. Curated evaluation data sets were provided for all three subtasks. These include test material obtained from the WMT25 General Machine Translation task as well as a collection of "challenge sets" that were developed by the organizers and members of the research community. A fourth subtask solicited the submission of these challenge sets.

The primary focus of this year's updated task

[1] www2.statmt.org/wmt25/mteval-subtask.html

Proceedings of the Tenth Conference on Machine Translation, pages 436–483
November 8-9, 2025

is on robust translation quality evaluation systems that can effectively detect translation errors generated by increasingly accurate LLM-based translators as well as handle previously unseen problems related to using LLMs for translation, such as output verbosity and incorrect output language. Newly, the evaluation data, originating from the General MT task, were intentionally chosen to be challenging for MT: they feature longer length, sourcing from some originally non-textual modalities, and translation into a wider variety of languages than in previous years.

Task 1 this year was designed to evaluate both MT metrics and QE systems. Reference-based automatic metrics score MT output by comparing the translation with a reference translation generated by human translators, who are instructed to translate "from scratch" without post-editing from MT. Reference-free quality estimation (QE) systems score translations solely based on their adherence to the source language. We collectively refer to all of these systems as "auto-raters" throughout the rest of the paper. All auto-raters were evaluated based on their agreement with human ratings when scoring MT systems and human translations at the system and segment level. In Task 2, systems are evaluated on their ability to detect and accurately annotate the spans of translation errors by contrasting them with the human error annotations. For Task 3, systems that correct a given translation are evaluated based on the quality of their post-edits, rewarding effective improvements with minimal changes.

Below are some of the key details and changes implemented for this year's shared task:

- **Subtasks**: As illustrated in Table 1, we solicited participation in segment-level quality score prediction, span-level error detection, and quality-informed post-editing.

- **Language Pairs**: Based on the General MT shared task, this year covers 16 language pairs, many of which are novel: English→{Czech, Estonian, Icelandic, Egyptian Arabic, Bhojpuri, Maasai, Russian, Serbian Cyrillic, Ukrainian, Japanese, Chinese, Italian, Korean}, Czech→{German, Ukrainian}, and Japanese→Chinese. The domains are Literary (short story), News, Social, Speech (video transcripts), and Social.

- **Human Evaluation**: Human evaluation was done as part of the General MT task, and these same annotations were then reused for our shared task. For 14 of the language pairs, annotations were conducted with the ESA protocol (Kocmi et al., 2024); the remaining two language pairs were annotated using the MQM protocol (Freitag et al., 2021). ESA annotations included two sets of human annotations per translation.

Task	Level	Meta-metric
1	system	SPA
1	segment	acc^*_{eq}
2	char	F1
3	segment	Δ-COMET

Table 2: Each language was given equal weight in the overall average.

- **Meta-Evaluation**: Each of the three subtasks employs its own individual meta-evaluation methods, described in more detail in the respective sections later in the paper. Task 1 follows the same approach as last year's Metrics task and uses two primary measures: soft pairwise accuracy (SPA) at the system level (Thompson et al., 2024), and "group-by-item" segment-level accuracy with tie calibration (acc^*_{eq}) at the segment level (Deutsch et al., 2023). Task 2 follows a similar approach to last year's QE task, and calculates the character-level F1 score between the predicted errors and the gold error spans, weighted to allow for half points for correctly identified spans with incorrect error severity. Task 3 evaluates the quality of the correction of the original MT using ΔCOMET as the primary measure and Gain-to-Edit Ratio (GER) to quantify editing efficiency.

- **MTME**: Similar to last year, all the data for Tasks 1 and 2 has been uploaded to the `mt-metrics-eval` codebase (MTME),[2] and all results in this paper are calculated with this analysis tool. We encourage developers of auto-rater systems to use MTME for greater reproducibility.

Our main findings are:

- The prediction of quality scores, with or without references, remains a challenge. Strong LLM-based auto-raters now top the rankings at the system level, where the task is to accurately identify the better MT system. At

[2] github.com/google-research/mt-metrics-eval

the segment level, LLM-based auto-raters still underperform; this year, unlike in recent years, reference-based baseline metrics (YiSi-1, chrF, and BERTScore) fill out the top three rank clusters, outranking recently strong trained metrics. We provide some analysis of this surprising result, but further analysis is needed to develop a better understanding of this outcome (Section 4).

- Auto-raters continue to struggle with precise error detection, span annotation, and severity classification, with significant gaps with human performance and large variations across language pairs (Section 5). This underscores the complexity of the task and highlights the critical need for advances in automated error analysis that better align with human judgments.

- While automatic translation correction systems can improve translation quality, this improvement is often at the cost of diverging from human-generated reference translations, indicating a gap between automated systems and human lexical choices, and that improvement does not necessarily mean alignment with human preferences (Section 6).

- By using carefully-crafted challenge sets, it is shown that current automatic MT evaluation systems still exhibit major weaknesses, including susceptibility to fluent but semantically irrelevant content, systematic gender bias, instability on low-quality or corner-case outputs, and poor correlation with human judgments for low-resource languages (Section 7).

The rest of the paper is organized as follows. Section 2 introduces our subtasks. Section 3 presents the evaluation set and the MT systems whose output was judged. Following that, the baselines, participants, meta-evaluation procedure, and main results of each subtask are discussed in Section 4 for segment-level quality score prediction, in Section 5 for span-level error annotation, and in Section 6 for quality-informed post-editing. Section 7 presents the submitted challenge sets, and Section 8 concludes.

2 Tasks

The shared task this year consisted of three primary subtasks that address translation quality assessment from three perspectives: (1) segment-level quality score prediction, (2) span-level translation error detection, and (3) quality-informed segment-level error correction. An additional fourth subtask solicited the submission of challenge sets that identify where automated metrics and auto-rater systems fail. All subtasks are introduced below:

2.1 Segment-Level Quality Score Prediction

The goal of the segment-level quality prediction subtask is to predict a quality score for each source–target segment pair in the evaluation set, with a reference translation optionally being provided. Depending on the language pair, the participants were asked to predict either the Error Span Annotation (ESA) score (Kocmi et al., 2024) or the Multi-dimensional Quality Metrics (MQM) score (Freitag et al., 2021). Submissions are evaluated and ranked based on their prediction correlations with these human-annotated scores at both the segment and system levels.

2.2 Span-Level Error Detection

In this subtask, the goal is to predict the precise span of each translation error along with its severity. For this subtask we use the error spans obtained from the MQM and ESA human annotations generated for the General MT primary task as the target "gold standard". Participants were asked to predict both the error spans (start and end indices) as well as the error severities (major or minor) within each segment. Submissions are evaluated and ranked based on their ability to correctly identify the presence of errors, correctly mark the spans of any identified errors, and correctly identify the severity of each of these errors.

2.3 Quality-Informed Segment-level Error Correction

The overarching goal of this subtask is to correct the output of machine translation. Recent work shows that joint optimization for QE and APE helps improve the performance of both tasks (Deoghare et al., 2023, 2024). Furthermore, fine-grained QE signals can also be leveraged to apply limited corrections and help mitigate over-correction, known as a common problem in APE systems (Deoghare et al., 2025). We invited participants to submit systems capable of automatically generating corrections for machine-translated text, given the source, the MT-generated target, and a QE-generated quality annotation of the MT. The objective is to ex-

plore how quality information, including both error span annotations and Direct Assessment (DA) scores, can inform automated error correction. For instance, sentence-level quality scores may help identify which segments require correction, while span-level annotations can be used for fine-grained, pinpointed corrections. Participants were provided with the quality information. Submissions were evaluated and ranked based on the quality of the corrections they generate, with as few changes as possible.

2.4 Challenge Sets

For the fourth year, our shared task included a subtask involving challenge sets. This subtask is inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017), which aimed at testing the generalizability of NLP systems beyond the distributions of their training data. Whereas the standard evaluation of the shared task is conducted on test sets containing generic text from real-world content, the challenge set evaluation is based on test sets designed with the aim of revealing the abilities or the weaknesses of the metrics or evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the subtask takes place in a decentralized manner: contrary to the main metric tasks, the test sets are not provided by the organizers but by different research teams, who are also responsible for analyzing and presenting the results (Section 7).

3 Evaluation Data

3.1 Data Sourcing and Translation

Similar to previous years' editions, the source sides, MT outputs, and reference texts for our shared task are mainly derived from the WMT25 General MT Shared Task (Kocmi et al., 2025a).

Newly this year, the source segments were automatically selected to be more challenging for translation systems, using a source-only difficulty estimator (Proietti et al., 2025). The test data domains cover news, literary (short stories), speech, and social. For the General MT shared task, some of this test material was multimodal: the speech data was provided as audio files with uncorrected automatic transcripts, and the social-media content was provided with screenshot images. However, for our shared task, we released only a text version of our evaluation set.

The selection of MT outputs was made based on an evaluation with automatic metrics (Kocmi et al., 2025b), giving slight prioritization to constrained (small, open-weight) systems. In keeping with our goal of exposing participants to a wide range of translation qualities and phenomena, we included output from around 20 different MT systems for each language pair. An exception is Task 3, which subsampled the original set for computational feasibility.

Reference translations were provided in 14 of our 16 language pairs, produced by professional translators from scratch. We ran English→Italian and English→Maasai as reference-free scenarios: the former because the General MT shared task intentionally did not produce any references, and the latter because the references produced by General MT were not yet available at the beginning of our evaluation period.

For more details regarding sourcing and translation of the test set, we refer the reader to the WMT25 General MT Shared Task (Kocmi et al., 2025a). All data has been released publicly.[3]

3.2 ESA and MQM Human Evaluation

This year, translations in most language pairs (English→{Czech, Estonian, Icelandic, Egyptian Arabic, Bhojpuri, Maasai, Russian, Serbian Cyrillic, Ukrainian, Japanese, Chinese, Italian} and Czech→{German, Ukrainian}) were evaluated with the ESA protocol (Kocmi et al., 2024). Japanese→Chinese and English→Korean were annotated with the MQM protocol (Freitag et al., 2021). The ESA protocol differs from MQM by not requiring the error categorization (fluency, punctuation, etc.) for each span and by providing a free-form 0% to 100% slider for assessing the final translation quality. Annotators were given guidance defining a score of 0% as "broken," 33% as "flawed," 66% as "good," and 100% as "perfect." In MQM, the translation score is instead derived mathematically from the count of error spans and their annotated severities. Generally, each minor error contributes –1 point while each major error counts as –5.

Our ESA annotations contain two judgments of each translation ("human1" and "human2"). This allows us to calculate intercoder agreement as a measure for task difficulty in each language and also provides a human "oracle" against which automated metrics can be compared (e.g. by treating

[3] github.com/wmt-conference/wmt25-general-mt

one human annotation as a "metric" and comparing it against actual auto-rater results). Additionally, the human annotations also contain control tasks (a fixed set of translations that all annotators annotated) designed to establish annotator reliability; this has been shown to work better than ad-hoc attention checks (Zouhar et al., 2025a).

In order to maximize differences between systems, human evaluation was limited to the top 50% of the most diversely translated inputs, computed with pairwise chrF between systems. Therefore, all source segments have either received two annotations for all selected systems, or none. This has the effect of avoiding spending human effort on annotating identical or similar translations, as well as translations that have little impact on the rankings of auto-rater systems (Zouhar et al., 2025b).

4 Task 1: Segment-Level Quality Score Prediction

This section presents the segment-level quality score prediction subtask in detail. We describe the auto-rater systems participating in the subtask in Section 4.1, our meta-evaluation procedure in Section 4.2, and our main results in Section 4.3. Section 4.4 reports some further analysis of the results beyond correlation and accuracy.

4.1 Participating Systems

We processed three distinct types of auto-rater systems participating in the segment-level quality score prediction subtask: baselines, official submissions, and "LLM as a judge" models. Each type is described in more detail below. A synthesized overview of all 48 systems is also given in Table 3, based on information provided by each participant at the time of submission. Full authoritative details are available in each team's separately prepared system description paper.

4.1.1 Baselines

We computed scores for several baseline systems in order to compare submissions against previous well-studied metrics.

SacreBLEU baselines We used the following metrics from SacreBLEU (Post, 2018):

- **BLEU (Papineni et al., 2002)** is based on the precision of n-grams between the MT output and its reference, weighted by a brevity penalty. We used the SacreBLEU command line with default arguments [4] for system-level BLEU, and we used the `-sl` argument to obtain segment-level BLEU.
- **spBLEU (NLLB-Team et al., 2022)** is the BLEU score computed with subword tokenization by the standardized FLORES-200 SentencePiece models. We used the SacreBLEU command line to compute system-level spBLEU,[5] and we used the `-sl` argument to obtain segment-level spBLEU.
- **chrF (Popović, 2015)** uses character n-grams instead of word n-grams to compare the MT output with the reference. We used the SacreBLEU command line with default arguments[6] for system-level chrF and used the `-sl` argument to obtain segment-level chrF.

BERTScore (Zhang et al., 2020) leverages contextual embeddings from pre-trained transformers to create soft alignments between words in hypothesis and reference segments using cosine similarity. Based on the alignment matrix, BERTScore returns a precision, recall, and F1 score. We used F1 without TF-IDF weighting.

COMET-22 (Rei et al., 2022a) is a learned metric fine-tuned using direct assessments from previous WMT translation shared tasks. This metric relies on sentence embeddings from the source, translation, and reference to produce a final score. We used the default model `wmt22-comet-da` provided in version 2.0.2 of the Unbabel/COMET framework. This model employs XLM-RoBERTa large as its backbone model and is trained on data from the 2017 to 2019 WMT shared tasks, in combination with the MLQE-PE corpus (Fomicheva et al., 2022).

COMETKIWI (Rei et al., 2022b) is a reference-free learned metric that functions similarly to BLEURT, but instead of encoding the translation along with its reference, it uses the source. We used the `wmt22-cometkiwi-da` model, which was a top-performing reference-free metric from the WMT 2022 shared task. This metric is fine-tuned on the same data as `wmt22-comet-da` using version 2.0.2 of the Unbabel/COMET framework.

[4] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|v:2.3.1. For into-Chinese, into-Japanese, and into-Korean language pairs, we used tok:zh, tok:ja-mecab, and tok:ko-mecab as the tokenizer, respectively.

[5] nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp|v:2.3.1

[6] chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|sp.:no|v:2.3.1

Team	Auto-Rater Name	Purpose	Category	Backbone Model	Uses Ref?	Supervised?
Organizers	BERTSCORE	Baseline	embedding similarity	XLM-RoBERTa	Yes	No
Organizers	BLEU	Baseline	lexical overlap	—	Yes	No
Organizers	CHRF	Baseline	lexical overlap	—	Yes	No
Organizers	COMET22	Baseline	fine-tuned encoder	XLM-RoBERTa	Yes	Yes
Organizers	COMETKIWI22	Baseline	fine-tuned encoder	InfoXLM	No	Yes
Sentinel metrics	SENTINEL-CAND	Baseline	fine-tuned encoder	XLM-RoBERTa	No	Yes
Sentinel metrics	SENTINEL-SRC	Baseline	fine-tuned encoder	XLM-RoBERTa	No	Yes
Organizers	SPBLEU	Baseline	lexical overlap	—	Yes	No
Organizers	YISI-1	Baseline	embedding similarity	XLM-RoBERTa, BERT-zh	Yes	No
Phrase	ENSEMBLESLICK	Primary	fine-tuned LLM	GPT variants	No	Yes
Microsoft Translator	GEMBA-V2	Primary	LLM-based	GPT 4.1 mini	No	No
hw-tsc	HW-TSC	Primary	?	?	No	No
MetricX-25	METRICX-25	Primary	fine-tuned LLM	Gemma 3 12B	No?	Yes
CUNI	MR7.2.1	Primary	fine-tuned LLM	Gemma 3 27B IT	No	Yes
KIT-ETH-UMich	POLYCAND-2	Primary	fine-tuned encoder	XLM-RoBERTa	No	Yes
Sujal_and_Astha	RANKEDCOMET	Primary	fine-tuned encoder	XLM-RoBERTa	Yes?	Yes
DCU_ADAPT	ROBERTA-LS	Primary	fine-tuned encoder	XLM-RoBERTa	Yes	Yes
Nvidia-Nemo	SEGALE-QE	Primary	fine-tuned LLM	Gemma 3 12B	No	Yes
TASER	TASER-NO-REF	Primary	LLM-based	OpenAI o3	No	No
UvA-MT	UVA-MT	Primary	LLM-based	Gemma 3 12B	No	No?
Phrase	AUTOLQA	Secondary	fine-tuned LLM	GPT variants	No?	Yes
Sujal_and_Astha	BASECOMET	Secondary	fine-tuned encoder	XLM-RoBERTa	Yes?	Yes
CUNI	COLLABPLUS	Secondary	ensemble	—	No?	Yes
Phrase	COLLABSLICK	Secondary	fine-tuned LLM	GPT variants	No?	Yes
hw-tsc	HW-TSC-BASE	Secondary	?	?	No?	No
hw-tsc	HW-TSC-MAX	Secondary	?	?	No?	No
DCU_ADAPT	LONG-CONTEXT	Secondary	fine-tuned	?	Yes?	Yes
MetricX-25	METRICX-25-QE	Secondary	fine-tuned LLM	Gemma 3 12B	No	Yes
MetricX-25	METRICX-25-REF	Secondary	fine-tuned LLM	Gemma 3 12B	Yes	Yes
CUNI	MR6	Secondary	fine-tuned LLM	Gemma 3 27B IT	No?	Yes
KIT-ETH-UMich	POLYCAND-1	Secondary	fine-tuned encoder	XLM-RoBERTa	No	Yes
KIT-ETH-UMich	POLYIC-3	Secondary	fine-tuned encoder	XLM-RoBERTa	No	Yes
Nvidia-Nemo	Q_MQM	Secondary	LLM-based	Qwen 3	No?	No
Nvidia-Nemo	Q_RELATIVE-MQM	Secondary	LLM-based	Qwen 3	No	No
DCU_ADAPT	ROBERTA-MULTI	Secondary	fine-tuned encoder	XLM-RoBERTa	Yes?	Yes
TASER	TASER-REF	Secondary	LLM-based	OpenAI o3	Yes?	No
Organizers	AYAEXPANSE-32B	LLM	LLM	AyaExpanse 32B	No	No
Organizers	AYAEXPANSE-8B	LLM	LLM	AyaExpanse 8B	No	No
Organizers	CLAUDE-4	LLM	LLM	Claude 4	No	No
Organizers	COMMANDA	LLM	LLM	CommandA	No	No
Organizers	COMMANDR7B	LLM	LLM	CommandR 7B	No	No
Organizers	DEEPSEEK-V3	LLM	LLM	DeepSeek V3	No	No
Organizers	GPT-4.1	LLM	LLM	GPT 4.1	No	No
Organizers	LLAMA-3.1-8B	LLM	LLM	Llama 3.1 8B	No	No
Organizers	LLAMA-4-MAVERICK	LLM	LLM	Llama 4 Maverick	No	No
Organizers	MISTRAL-7B	LLM	LLM	Mistral 7B	No	No
Organizers	QWEN2.5-7B	LLM	LLM	Qwen 2.5 7B	No	No
Organizers	QWEN3-235B	LLM	LLM	Qwen 3 235B	No	No

Table 3: Summary of Task 1 participants. We distinguish four different purposes of participation: as a baseline, as an official primary submission, as an official secondary submission, or as an "LLM as a judge." Basic self-submitted properties of each entrant are summarized above, sorted by auto-rater name.

Sentinel baselines We also included two metrics from the Sentinel family. Unlike the other baselines, Sentinel metrics are intentionally formulated to lack important information when assigning their scores. They are instead meant as a probing mechanism to highlight evaluation scenarios that may be "too easy" or that are prone to spurious correlations, if the Sentinel metrics place competitively among other evaluators that have access to more complete information.

- **SENTINEL-SRC-25 (Proietti et al., 2025)** predicts the quality of a translation solely based on its source string, without considering the reference or even the translation itself. It is an updated version of the original SENTINEL-SRC: a regression model based on XLM-RoBERTa, trained with data from previous WMT editions up through and including the WMT 2024 test set.

- **SENTINEL-CAND (Perrella et al., 2024)** assesses the quality of a translation based on the output string alone, without taking the source or reference into account. It is also based on XLM-RoBERTa, trained with WMT data up

through 2022.

YISI-1 (Lo, 2019) is an MT evaluation metric that measures the semantic similarity between a machine translation and human references by aggregating the IDF-weighted lexical semantic similarities based on the contextual embeddings extracted from pre-trained language models (BERT-base-chinese for evaluating Chinese and XLM-RoBERTa for evaluating other target languages in this shared task).

4.1.2 Official Submissions

Each team participating in Task 1 was allowed to submit one primary and up to two secondary systems for meta-evaluation. The primary systems are described below. Secondary systems are included in the general tabular overview (Table 3).

ENSEMBLESLICK (Hrabal et al., 2025) For Task 1, this system uses a combination of Phrase proprietary fine-tuned GTE and similar models and fine-tuned GPT-4o-mini.

GEMBA-V2 (Junczys-Dowmunt, 2025) GEMBA-V2 is an updated version of GEMBA (Kocmi and Federmann, 2023).

HW-TSC (Luo et al., 2025) This system's approach integrates sentence segmentation tools and dynamic programming to construct sentence-level alignments between source and translated texts, then adapts sentence-level evaluation models to document-level assessment via sliding-window aggregation.

METRICX-25 (Juraska et al., 2025) METRICX-25 is an encoder-only regression model initialized from Gemma 3 12B and fine-tuned on publicly available DA and MQM scores from WMT 2015–23 in a two-stage fashion. Similar to METRICX-24, the first stage uses z-normalized DA scores, and the second stage uses a mixture of raw DA scores (rescaled to the MQM range of 0–25) and MQM scores. Due to the dual nature of meta-evaluation this year (ESA/DA vs. MQM), a score type indication is included in the input, indicating for each training example whether it corresponds to a DA or MQM score.

MR7.2.1 (Hrabal et al., 2025) This submission experimented with the Gemma 3 27B IT model prompted using the DSPy framework and using its MIPROv2 optimizer. The system first generates seven 0–10 integer scores for various aspects of the translation (e.g. "accuracy and completeness" or "fluency and coherence"). Afterwards, it generates the overall 0–100 score.

POLYCAND-2 (Züfle et al., 2025) The supervised reference-less metric $\text{COMET}_{\text{poly-}*}$ has similar architecture and training to standard COMET but incorporates additional information beyond one single translation. $\text{COMET}_{\text{poly-cand2}}$ incorporates two alternative translations of the same source segment (provided by other translation systems) to better contextualize and assess the quality of the translation being scored. The metric was trained on a limited combination of DA, ESA, and MQM data on a unified scale.

RANKEDCOMET (Maharjan and Shrestha, 2025) This system is based on the pre-trained Unbabel/wmt22-comet-da model, deployed in a zero-shot inference setting. Raw segment-level quality scores are generated and then post-processed with per-language-pair rank normalization. This method transforms raw scores into a calibrated distribution that significantly improved correlation with the preliminary evaluation metrics.

ROBERTA-LS (Haq and Osuji, 2025) ROBERTA-LS (Roberta Long-Span) is a reference-based evaluation metric built using the COMET framework. Designed to provide multi-sentence quality scores, it is trained on augmented long-context data that captures translation quality beyond isolated sentences. To construct the long-span MT evaluation dataset, adjacent short segments are concatenated, and a multi-segment quality score is computed as a length-weighted average of their original scores. Unbabel/wmt22-comet-da and XLM-RoBERTa-base are fine-tuned on the augmented data.

SEGALE-QE (Yan et al., 2025) This system extends METRICX to long texts by adding a pipeline before running the metric. It first segments the data down to individual sentences with Ersatz, then runs Vecalign to align system translations to the source. Vecalign's deletion penalty is adaptively adjusted to obtain good alignments that exclude over/under-translation to the maximum extent possible. When over/under-translated sentences are identified, they are assigned a score of 25. Individual sentence scores are then averaged to form the score for the long-form translation pair.

TASER-NO-REF (Maheswaran et al., 2025) TASER (Translation Assessment via Systematic Evaluation and Reasoning) is a Large Reasoning Model–based metric for translation quality assessment. This metric uses OpenAI's o3 to estimate the quality of a translation in reference-free scenarios. It posits that LRMs are capable of better assessing the quality of translations than vanilla LLMs with advanced prompting strategies.

UVA-MT (Wu and Monz, 2025) This system calibrates quality estimation and likelihood on the google/gemma3-12b-it model, then directly uses the token average likelihood as a metric for quality estimation. No human annotation data is used; the only reliance is on a translation's likelihood as the metric. The same resulting model was also submitted to the WMT 2025 General Translation Task, meaning that it grades its own output as part of the segment-level quality prediction task.

4.1.3 LLMs as Judges

As a third category of system, the shared task organizers obtained quality scores on our test set from 12 different publicly available large language models, using their standard APIs and a templated prompt. These submissions test the ability of general-purpose LLMs as judges of translation quality without fine-tuning or few-shot examples.

LLMs for which we obtained quality scores are: AyaExpanse 8B, AyaExpanse 32B, Claude 4, Command A, Command R7B, DeepSeek V3, GPT 4.1, Llama 3.1 8B, Llama 4 Maverick, Mistral 7B, Qwen 2.5 7B, and Qwen 3 235B. The templated ESA-like prompt is given in Appendix A.

4.2 Meta-Evaluation

The goal of auto-rater meta-evaluation is to quantify how well automatic systems agree with human ratings of translation quality. There are a multitude of ways to approach this problem, as evidenced by the variety of solutions proposed by previous years' editions of the shared task.[7] Ranking-based approaches (traditionally Spearman's ρ, Kendall's τ, or pairwise accuracy) assume the least about the relative shapes of the score distributions: only the directionality matters, and the magnitude of difference is ignored. Linear correlation (traditionally Pearson's r) captures magnitude but thereby assumes a constant slope to the scores and can be unduly influenced by outliers (Mathur et al., 2020).

We follow the same approach as last year to this year's meta-evaluation of Task 1, focusing on improved ranking-based methods.

At the system level, we use soft pairwise accuracy, or SPA (Thompson et al., 2024). SPA uses p-values as a proxy for certainty about the difference between two systems, calculated over both the auto-rater and human scores. This rewards auto-raters that result in the same statistical conclusion as the human scores. However, computation of the p-values requires repeated resampling of segments in order to determine the statistical range of system-level scores. For efficiency of meta-evaluation, SPA *averages* the segment-level scores in each resample as a proxy for the system-level score. Such averaging is a technically incorrect aggregation method for BLEU, CHRF, and a number of other submissions that self-reported that they employ some more complicated methodology.

At the segment level, we again follow last year's process and meta-evaluate metrics using "group-by-item" segment-level accuracy with tie calibration (Deutsch et al., 2023), denoted acc^*_{eq}. Group-by-item processing, recommended by Perrella et al. (2024) avoids pairwise comparisons between translations originating from different source segments.

Because SPA and acc^*_{eq}meta-metrics are based on ranking, we do not perform any normalization on the raw scores output by the auto-raters or annotated by the humans.

We assign ranks to auto-raters based on their significance clusters in the same way that we did last year. We compare all pairs of auto-raters and determine whether the difference in their correlation scores is significant according to the PERM-BOTH hypothesis test of Deutsch et al. (2021). We use 1000 re-sampling runs and set $p = 0.05$. As advocated by Wei et al. (2022), we divide the sample into blocks of 100, compute significance after each block (cumulative over all blocks sampled so far), and stop early if the p-value is < 0.02 or > 0.50. To calculate p-values for SPA, we use a paired permutation test (Noreen, 1989) with 1000 resamples.

Given the significance results (p-values) for all pairs of auto-raters, ranks are assigned starting with the highest-scoring auto-rater. We move down the list of auto-raters in descending order by score, assigning rank 1 to all auto-raters until we encounter the first one that is significantly different from any that have been visited so far. That auto-rater is

[7] See Section 5 of Thompson et al. (2024) and Table 1 of Deutsch et al. (2023) for nice summaries of the approaches taken in prior WMT shared tasks to meta-evaluation at, respectively, the system and segment level.

assigned rank 2, and the process is repeated. This continues until all auto-raters have been assigned a rank. Note that this is a greedy algorithm, and hence it can place two auto-raters that are statistically indistinguishable in different clusters.

The code for running the meta-evaluation is available in the `mt-metrics-eval` library.[8]

While the segments in language pairs evaluated with ESA received two independent human judgments (as per Section 3.2), most of the results and analyses presented in Section 4.3 and Section 4.4 are based on the first complete annotation that we received, which we refer to as "human1." The second set of human judgements ("human2") did not arrive in time to permit a complete analysis with the `mt-metrics-eval` package.

4.3 Main Results

Summarized results for the quality score prediction subtask are shown in Table 4 and Table 5. Table 4 reports average performance across the 14 language pairs for which references were provided, while Table 5 covers the remaining two reference-less language pairs. (Since we do not have complete information about which auto-raters make use of the reference or may do so optionally, we list in Table 5 all the entrants that submitted scores for reference-free language pairs.) Note that three participating systems (ROBERTA-LS, LONG-CONTEXT, and ROBERTA-MULTI) returned output for only three language pairs and are thus not included in either summary table for lack of a fair comparison. Full detailed results broken down by individual language pair are given in Table 19 (part 1) and Table 20 (part 2) in Appendix B; all participating systems appear there.[9]

A striking pattern in this year's results is the strong performance of many baseline systems — especially those based on lexical overlap or embedding similarity. YISI-1, CHRF, and BERTSCORE fill out the top three rank clusters when judged at the segment level in the presence of references. General-purpose LLMs do quite poorly in terms of correlation with human judgments at the segment level, but their system-level performance is better, led by GPT 4.1 and Claude 4. GPT 4.1 and TASER-REF, a reasoning-based model, achieve top-tier performance on system-level correlation when a reference is present; reference-free models are led by GEMBA-V2.[10] Sentinel models, as desired, rank lowly throughout.

	Avg All		Avg Sys		Avg Seg	
	Rank	Corr	Rank	Corr	Rank	Corr
Baselines						
YiSi-1	2	0.674	4	0.791	1	0.558
chrF	2	0.672	4	0.789	2	0.554
spBLEU	3	0.668	5	0.784	4	0.551
BERTScore	4	0.662	6	0.770	3	0.553
BLEU	5	0.657	6	0.770	6	0.543
COMET22	8	0.624	9	0.709	8	0.539
sentinel-cand	17	0.533	16	0.572	17	0.494
COMETKiwi22	19	0.505	18	0.526	20	0.484
sentinel-src	25	0.351	19	0.509	37	0.193
Primary						
GEMBA-v2	2	0.672	3	0.811	9	0.533
TASER-No-Ref	3	0.666	2	0.833	16	0.499
rankedCOMET	6	0.627	8	0.716	8	0.539
MetricX-25	8	0.621	9	0.711	10	0.530
mr7_2_1	9	0.614	6	0.760	24	0.467
SEGALE-QE	13	0.581	12	0.654	12	0.509
Polycand-2	14	0.566	13	0.626	13	0.506
Q_Relative-MQM	15	0.564	7	0.737	28	0.391
EnsembleSlick	17	0.539	15	0.600	23	0.478
hw-tsc	18	0.524	17	0.557	18	0.490
UvA-MT	21	0.465	20	0.466	25	0.464
Secondary						
TASER-Ref	1	0.698	1	0.846	5	0.549
MetricX-25-Ref	6	0.633	8	0.727	7	0.539
baseCOMET	7	0.624	10	0.709	7	0.539
MetricX-25-QE	10	0.602	11	0.681	11	0.524
mr6	10	0.598	7	0.738	26	0.458
Q_MQM	14	0.568	7	0.736	27	0.399
Polyic-3	16	0.555	14	0.607	14	0.503
AutoLQA	16	0.553	10	0.707	27	0.398
Polycand-1	16	0.554	14	0.606	15	0.501
CollabPlus	16	0.548	13	0.612	20	0.485
CollabSlick	16	0.548	14	0.609	19	0.487
hw-tsc-max	19	0.509	18	0.536	21	0.483
hw-tsc-base	20	0.499	19	0.518	22	0.479
LLM-as-a-judge						
GPT-4_1	3	0.669	1	0.849	18	0.489
CommandA	11	0.597	3	0.812	29	0.382
Claude-4	12	0.593	2	0.833	31	0.352
DeepSeek-V3	13	0.582	4	0.797	30	0.368
Qwen3-235B	13	0.579	4	0.790	30	0.368
Qwen2_5-7B	19	0.507	12	0.667	32	0.347
AyaExpanse-32B	19	0.500	7	0.732	34	0.269
Llama-3_1-8B	21	0.466	12	0.663	34	0.269
Llama-4-Maverick	22	0.453	7	0.730	38	0.176
CommandR7B	23	0.408	16	0.568	35	0.248
Mistral-7B	23	0.401	18	0.527	33	0.274
AyaExpanse-8B	24	0.387	15	0.576	36	0.199

Table 4: Task 1 results summary against the "human1" annotation for all language pairs with references.

[8] github.com/google-research/mt-metrics-eval

[9] We also show in Appendix B the results on the ESA language pairs using the "human2" annotation as the gold standard.

[10] TASER-REF also ranks first or second in reference-free evaluation; even though it is labeled as a reference-using metric, it was submitted with scores for the segments without references as well.

	Avg All		Avg Sys		Avg Seg	
	Rank	Corr	Rank	Corr	Rank	Corr
Baselines						
sentinel-cand	8	0.542	6	0.593	7	0.492
COMETKiwi22	10	0.501	9	0.517	8	0.485
sentinel-src	12	0.417	9	0.501	23	0.333
Primary						
GEMBA-v2	1	0.638	1	0.764	2	0.512
TASER-No-Ref	3	0.601	3	0.710	7	0.493
mr7_2_1	4	0.581	3	0.702	11	0.460
SEGALE-QE	4	0.570	5	0.632	3	0.508
EnsembleSlick	7	0.550	5	0.609	7	0.491
Q_Relative-MQM	7	0.549	4	0.686	18	0.413
MetricX-25	7	0.548	7	0.583	2	0.514
Polycand-2	7	0.547	6	0.599	5	0.495
rankedCOMET	8	0.542	7	0.592	7	0.493
hw-tsc	8	0.538	7	0.584	7	0.491
UvA-MT	11	0.485	10	0.494	10	0.476
Secondary						
TASER-Ref	1	0.633	2	0.738	1	0.528
Q_MQM	4	0.578	2	0.740	17	0.415
mr6	5	0.569	4	0.682	12	0.456
MetricX-25-QE	6	0.554	6	0.594	2	0.514
CollabSlick	6	0.553	5	0.609	4	0.498
baseCOMET	7	0.549	6	0.605	6	0.493
Polycand-1	8	0.536	7	0.579	7	0.493
Polyic-3	9	0.532	8	0.570	5	0.495
CollabPlus	9	0.528	8	0.557	4	0.498
AutoLQA	9	0.525	6	0.597	12	0.453
hw-tsc-max	10	0.512	9	0.541	9	0.483
hw-tsc-base	10	0.512	9	0.541	9	0.483
LLM-as-a-judge						
GPT-4_1	2	0.611	2	0.725	4	0.496
CommandA	4	0.577	3	0.711	13	0.443
Claude-4	4	0.574	2	0.729	16	0.419
Qwen3-235B	4	0.573	3	0.712	14	0.434
DeepSeek-V3	5	0.565	3	0.716	17	0.413
Qwen2_5-7B	5	0.562	3	0.696	15	0.428
AyaExpanse-32B	7	0.543	3	0.702	19	0.383
CommandR7B	9	0.529	4	0.685	20	0.373
Llama-3_1-8B	10	0.513	4	0.662	21	0.364
Llama-4-Maverick	11	0.485	4	0.669	24	0.301
Mistral-7B	11	0.484	6	0.594	20	0.373
AyaExpanse-8B	11	0.473	5	0.609	22	0.337

Table 5: Task 1 results summary against the "human1" annotations for all language pairs without references (i.e. English–Italian and English–Maasai).

Divergence in these results compared to recent years may be due to a variety of causes. In particular, we note that the source texts were intentionally chosen to be difficult, that they consist of longer paragraph-like segments, and that there are therefore fewer segments available for scoring in each language pair than in the past. Further, we ran the quality score prediction task in a wider variety of language pairs, with a correspondingly wider variety in the quality of the MT output being judged.

We will further explore a few interesting facets of the results in the following section.

4.4 Analysis

In this section, we discuss the performance of MT auto-rater systems from several additional perspectives, in order to interpret our results, to provide further insights on strength and weakness of various classes of auto-raters, and to shed light on upcoming challenges in automated translation quality evaluation research.

4.4.1 SPA vs. Pairwise Accuracy

Because the results in Section 4.3 differ from those obtained in other recent years — notably on the relative strength of string-based metrics — we ran a contrastive evaluation where the system-level meta-metric was changed from SPA to "hard" pairwise accuracy instead. System-level pairwise accuracy (Kocmi et al., 2021) was used as a meta-evaluation metric in the WMT metrics task from 2021 through 2023. This method, similarly to SPA, compares MT system pair ranking decisions between humans and auto-raters, but it ignores the magnitude of the human and auto-rater score differences. Pairwise accuracy also does not require any resampling or averaging of segment-level scores to create proxies for system-level scores.

Table 6 shows the correlation results obtained by using each system-level meta-metric, for language pairs that were provided with references, against the "human1" annotations as the gold standard. The "SPA" columns of the table are equivalent to the system-level data shown in Table 4, repeated here for easy side-by-side comparison. The "Pair Acc" columns show the analogous results using pairwise accuracy. (Note that the use of pairwise accuracy leads to a smaller number of statistically significant auto-rater clusters, so the rank ordinals are not comparable between the "SPA" and "Pair Acc" columns.) The right-most "Diff" column shows the differences in correlation between the two settings.

SPA and pairwise accuracy produced very similar correlation scores and overall rankings of auto-raters. Our contrastive experiments therefore did not shed light on why string-based metrics are performing better than expected. The main difference we observe is that SPA produced substantially more statistically significant comparisons, resulting in twice as many (20 vs. 10) significance clusters. This finding is consistent with Thompson et al. (2024).

	SPA		Pair Acc		Diff
	Rank	Corr	Rank	Corr	ΔCorr
Baselines					
YiSi-1	4	0.791	3	0.795	0.004
chrF	4	0.789	3	0.783	−0.006
spBLEU	5	0.784	3	0.780	−0.004
BERTScore	6	0.770	3	0.772	0.002
BLEU	6	0.770	3	0.767	−0.003
COMET22	9	0.709	5	0.701	−0.008
sentinel-cand	16	0.572	8	0.573	0.001
COMETKiwi22	18	0.526	9	0.515	−0.011
sentinel-src	19	0.509	9	0.476	−0.033
Primary					
TASER-No-Ref	2	0.833	2	0.828	−0.005
GEMBA-v2	3	0.811	2	0.815	0.004
mr7_2_1	6	0.760	4	0.760	0.000
MetricX-25	9	0.711	5	0.705	−0.006
rankedCOMET	8	0.716	5	0.703	−0.013
SEGALE-QE	12	0.654	6	0.654	0.000
Polycand-2	13	0.626	6	0.627	0.001
EnsembleSlick	15	0.600	8	0.597	−0.003
hw-tsc	17	0.557	8	0.555	−0.002
UvA-MT	20	0.466	10	0.468	0.002
Secondary					
TASER-Ref	1	0.846	1	0.846	0.000
Q_MQM	7	0.736	4	0.742	0.006
mr6	7	0.738	4	0.741	0.003
Q_Relative-MQM	7	0.737	4	0.740	0.003
MetricX-25-Ref	8	0.727	4	0.720	−0.007
AutoLQA	10	0.707	5	0.708	0.001
baseCOMET	10	0.709	5	0.702	−0.007
MetricX-25-QE	11	0.681	6	0.673	−0.008
CollabSlick	14	0.609	7	0.614	0.005
CollabPlus	13	0.612	7	0.613	0.001
Polycand-1	14	0.606	7	0.608	0.002
Polyic-3	14	0.607	7	0.604	−0.003
hw-tsc-max	18	0.536	9	0.538	0.002
hw-tsc-base	19	0.518	9	0.524	0.006
LLM-as-a-judge					
GPT-4_1	1	0.849	1	0.849	0.000
Claude-4	2	0.833	1	0.839	0.006
CommandA	3	0.812	2	0.813	0.001
DeepSeek-V3	4	0.797	3	0.802	0.005
Qwen3-235B	4	0.790	3	0.794	0.004
Llama-4-Maverick	7	0.730	4	0.741	0.011
AyaExpanse-32B	7	0.732	4	0.734	0.002
Qwen2_5-7B	12	0.667	6	0.669	0.002
Llama-3_1-8B	12	0.663	6	0.660	−0.003
CommandR7B	16	0.568	8	0.575	0.007
AyaExpanse-8B	15	0.576	8	0.550	−0.026
Mistral-7B	18	0.527	9	0.507	−0.020

Table 6: System-level correlation using either SPA (equivalent to Table 4) or pairwise accuracy, against the "human1" annotation for all language pairs with references. The scores and overall rankings are quite similar for SPA and pairwise accuracy, but SPA produces substantially more (20 vs. 10) significance clusters.

	All MT		Select MT		Diff
	Rank	Corr	Rank	Corr	ΔCorr
Baselines					
spBLEU	5	0.784	3	0.789	0.006
YiSi-1	4	0.791	3	0.789	−0.002
chrF	4	0.789	3	0.786	−0.003
COMET22	9	0.709	4	0.770	0.060
BLEU	6	0.770	4	0.769	−0.001
BERTScore	6	0.770	4	0.764	−0.006
sentinel-cand	16	0.572	7	0.684	0.112
COMETKiwi22	18	0.526	10	0.551	0.025
sentinel-src	19	0.509	11	0.495	−0.014
Primary					
TASER-No-Ref	2	0.833	1	0.836	0.003
GEMBA-v2	3	0.811	2	0.819	0.008
rankedCOMET	8	0.716	4	0.775	0.059
mr7_2_1	6	0.760	4	0.761	0.001
MetricX-25	9	0.711	5	0.756	0.045
Q_Relative-MQM	7	0.737	6	0.719	−0.017
SEGALE-QE	12	0.654	6	0.693	0.039
Polycand-2	13	0.626	7	0.687	0.062
EnsembleSlick	15	0.600	7	0.669	0.069
hw-tsc	17	0.557	8	0.652	0.095
UvA-MT	20	0.466	12	0.374	−0.091
Secondary					
TASER-Ref	1	0.846	1	0.840	−0.006
baseCOMET	10	0.709	4	0.770	0.061
MetricX-25-Ref	8	0.727	5	0.749	0.022
MetricX-25-QE	11	0.681	5	0.747	0.066
AutoLQA	10	0.707	5	0.743	0.036
mr6	7	0.738	5	0.740	0.002
Q_MQM	7	0.736	6	0.719	−0.017
Polyic-3	14	0.607	6	0.699	0.092
Polycand-1	14	0.606	7	0.684	0.078
CollabPlus	13	0.612	7	0.676	0.064
CollabSlick	14	0.609	7	0.671	0.061
hw-tsc-max	18	0.536	9	0.602	0.067
hw-tsc-base	19	0.518	10	0.573	0.055
LLM-as-a-judge					
GPT-4_1	1	0.849	1	0.840	−0.009
Claude-4	2	0.833	1	0.834	0.002
CommandA	3	0.812	2	0.820	0.008
DeepSeek-V3	4	0.797	2	0.814	0.017
Qwen3-235B	4	0.790	3	0.796	0.006
AyaExpanse-32B	7	0.732	5	0.750	0.019
Llama-4-Maverick	7	0.730	6	0.723	−0.002
Llama-3_1-8B	12	0.663	8	0.633	−0.030
Qwen2_5-7B	12	0.667	8	0.631	−0.035
AyaExpanse-8B	15	0.576	10	0.530	−0.045
Mistral-7B	18	0.527	10	0.524	−0.004
CommandR7B	16	0.568	11	0.478	−0.090

Table 7: Average system-level correlations using either all available MT output (equal to Table 4) or only output from selected MT systems unlikely to have hill-climbed on metrics. Correlations are computed against "human1" annotations.

4.4.2 MT Systems That Hill-Climb Metrics

Some systems in the WMT General MT task (Kocmi et al., 2025a), whose output we rely on to create this task's test set, are tuned against automatic metrics (for example, as part of the reward model). This presents a challenge when the same automatic metric is now asked to judge the qual-

ity of the MT: a bias towards the translations that have been specially optimized towards it could negatively affect the metric's correlation with independent human judgments.

To investigate this effect, we conducted a follow-up analysis in which we calculated auto-rater performance only when judging output from a population of MT systems that are highly unlikely to be metric-tuned. These selected MT systems primarily consist of general-purpose LLMs and publicly available MT services. We repeated the system-level meta-evaluation from Section 4.2, using SPA, on this reduced portion of our test set.

Table 7 shows a comparison of the results in the two cases. The "All MT" column represents the rank and average system-level correlation of each participant for the 14 language pairs in the original test set that provide reference translations. (This section of the table is equivalent to the system-level columns of Table 4.) In the "Select MT" column, we display the analogous results on the smaller test set. (Note that the smaller test set leads to a smaller number of statistically significant metric clusters, so the rank ordinals are not comparable between the two columns.) The right-most "Diff" column shows the differences in average correlation between the two settings.

As expected, metrics that are the most likely targets of MT hill-climbing see their correlations with human judgments improve once we remove the affected MT systems. This is true most visibly for the numerous variations of COMET — including all three POLYCAND* auto-raters, all of which rank in the top 10 "most improved." Conversely, the smallest changes in average correlation tend to come from classic string- or embedding-based metrics, which are unlikely to serve as modern-day MT optimization targets, as well as TASER variants and high-performing general-purpose LLMs.

With these results in mind, we caution MT practitioners against evaluating system variants according to the same metrics that played any role in the systems' training process.

4.4.3 Detecting Catastrophic Translations

The distribution of translation quality varies greatly across languages. For example, high-resource languages in our test set tend to come with the most translations that are near perfect, while even state-of-the-art MT systems struggle with lower-resource languages or languages and domains not previously in WMT. We show this distribution of human-annotated ESA scores in Figure 1. This is a problem for trained auto-rater systems, which underperform in unseen domains and tend to follow the language distribution; they are thus likely to score translations into a low-resource language as lower in quality and they have greater variance (Zouhar et al., 2024a,b).

This year's language pairs created a new issue for MT systems: incorrect output language. Specifically, for some language pairs, some MT systems outputted the wrong language, dialect, or script.

Auto-Rater	en-ar	en-bho	en-sr	en-et	en-is	en-ru
Human	98%	78%	86%	24%	46%	13%
Claude-4	77%	61%	73%	19%	49%	16%
GPT-4	84%	39%	74%	23%	54%	15%
TASER-Ref	77%	50%	56%	21%	53%	16%
COMETKiwi22	84%	58%	55%	14%	41%	13%
Polyic-3	81%	67%	33%	15%	48%	12%
UvA-MT	76%	49%	69%	16%	36%	10%
Polycand-2	80%	48%	48%	16%	48%	12%
MetricX-25-Ref	79%	47%	51%	17%	44%	13%
DeepSeek-V3	76%	28%	69%	16%	45%	14%
MetricX-25-QE	77%	45%	51%	17%	45%	13%
BERTScore	84%	51%	50%	16%	36%	10%
Polycand-1	80%	53%	39%	14%	48%	12%
hw-tsc	78%	56%	48%	13%	37%	12%
SEGALE-QE	76%	49%	43%	17%	46%	12%
hw-tsc-base	79%	56%	46%	13%	35%	12%
YiSi-1	79%	43%	47%	18%	43%	11%
hw-tsc-max	77%	56%	46%	14%	35%	12%
CommandA	77%	24%	67%	16%	38%	14%
MetricX-25	77%	37%	49%	15%	42%	13%
sentinel-cand	80%	54%	32%	13%	41%	11%
COMET22	79%	30%	42%	18%	44%	12%
rankedCOMET	79%	30%	42%	18%	44%	12%
baseCOMET	79%	30%	42%	18%	44%	12%
mr7_2_1	78%	23%	53%	19%	37%	15%
Llama-4-Maverick	76%	25%	53%	13%	43%	13%
mr6	76%	27%	52%	16%	36%	13%
Qwen3-235B	76%	22%	50%	19%	35%	14%
Q_Relative-MQM	76%	23%	56%	12%	35%	14%
Q_MQM	76%	23%	45%	11%	35%	15%
GEMBA-v2	76%	25%	29%	13%	47%	14%
chrF	80%	36%	17%	17%	41%	9%
TASER-No-Ref	76%	36%	23%	9%	45%	12%
spBLEU	82%	34%	18%	18%	38%	10%
CollabSlick	76%	52%	8%	16%	35%	10%
CollabPlus	76%	51%	8%	14%	36%	11%
Qwen2	76%	32%	35%	14%	28%	11%
BLEU	80%	39%	18%	14%	35%	9%
EnsembleSlick	76%	50%	7%	16%	35%	10%
Mistral-7B	77%	31%	34%	9%	26%	13%
AyaExpanse-32B	76%	19%	40%	10%	29%	13%
Llama-3	76%	19%	29%	15%	31%	11%
AyaExpanse-8B	76%	21%	27%	11%	28%	13%
CommandR7B	77%	31%	24%	8%	27%	11%
AutoLQA	76%	19%	17%	11%	34%	11%
sentinel-src	76%	18%	8%	14%	26%	13%

Table 8: Ability of auto-raters to detect catastrophic translations (best threshold for F_1). Rows are ordered by average performance; auto-raters perform worse than human for languages with bimodal distributions (Figure 1).

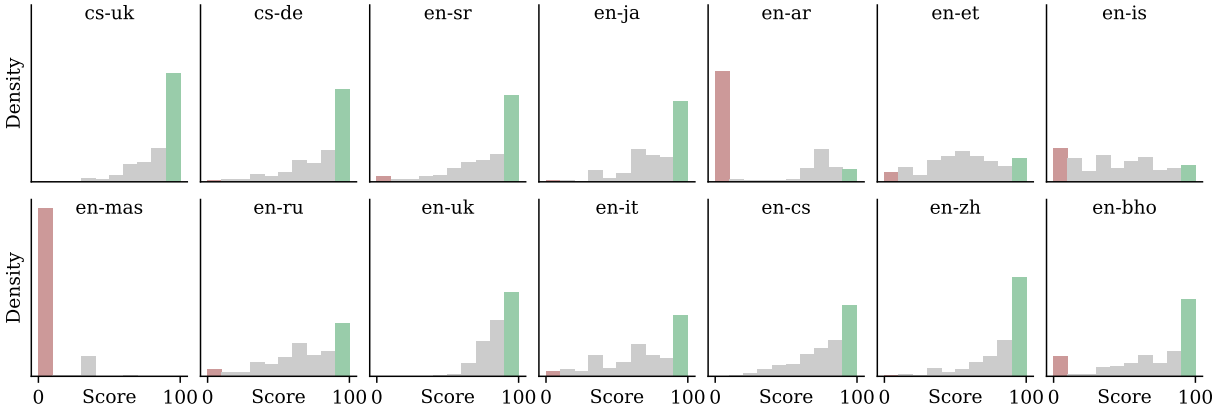


Figure 1: Human score distribution across languages evaluated with the ESA protocol. The pink bar corresponds to translation with less than 10 ESA points, and the green bar shows translations with above 90 ESA points.

Examples include English→Egyptian Arabic (oftentimes incorrectly translated as modern standard Arabic), English→Bhojpuri (incorrectly translated as a mixture of languages), or English→Serbian (incorrectly translated into a Latin script). These systems still made it into the human evaluation because the automatic evaluation filter used did not flag any major issues. In this section, we discuss the failure of auto-rater systems to detect catastrophic translation failure, in use-cases similar to those used by the General MT shared task.

We investigate select languages that do have a notable proportion of catastrophic translations, defined as those receiving an ESA score lower than 10. For those, we collect the set of catastrophic translations and select the threshold for each auto-rater to classify a catastrophic translation that maximizes the F_1 score (true positive = ESA score < 10). This threshold would, in theory, be possible to be used for identifying catastrophic translations without any human input. The results are shown in Table 8. For languages where the mismatch in languages, dialects, and scripts was a critical issue (Arabic, Bhojpuri, Serbian), the second human annotator was able to detect the catastrophic translations much better than automatic systems. This leaves headroom for improvements for auto-rater systems, which have to be able to score translations from state-of-the-art LLMs that might not always follow the instructions.

4.4.4 Utility of the Reference

Having investigated auto-rater systems capabilities in the face of especially poor MT output, we now turn to the complementary question: a study of our submissions' abilities to cope with poor *reference* translations.

Auto-Rater	**ΔCorr**	**Auto-Rater**	**ΔCorr**
YiSi-1	−0.163	CommandR7B	0.201
chrF	−0.152	Llama-3_1-8B	0.199
BLEU	−0.151	AyaExpanse-8B	0.189
BERTScore	−0.148	Qwen2_5-7B	0.178
spBLEU	−0.143	Mistral-7B	0.164
baseCOMET	−0.043	mr7_2_1	0.160
COMET22	−0.043	Q_MQM	0.144
hw-tsc-max	−0.036	Q_Relative-MQM	0.139
GPT-4_1	−0.030	AyaExpanse-32B	0.137
rankedCOMET	−0.030	mr6	0.112

Table 9: Auto-raters recording the largest drops (left) and gains (right) in average system-level correlation between language pairs with high- and low-quality references.

We would expect to see a divergence in performance between auto-raters that adhere closely to the reference and those that are reference-free. When the provided reference is itself of poor quality, auto-raters in the first group may be misled into misjudging the MT output. When the reference is quite accurate, on the other hand, auto-raters in the second group may suffer from not being able to consult it. Below we examine each of these cases individually.

For this analysis, we divide the language pairs in our test set into groups based on the *references'* performance in the human evaluation. In the WMT General Translation task (Kocmi et al., 2025a), the reference translation was judged to fall alone into the top-ranked cluster of "systems" in five language pairs: English→Arabic, English→Estonian, English→Icelandic, English→Japanese, and Japanese→English. Conversely, the reference placed relatively lowly in English→Russian (rank

9–11 of 19), English→Chinese (11–13 of 19), and Czech→German (9–12 of 21).

We extract system-level correlations for each participating auto-rater out of Table 19 and Table 20 in Appendix B in order to compute the average correlation separately per each group of language pairs. Since a strong auto-rater is more likely to outperform a weak auto-rater in *any* language pair, we compare instead the difference in correlation for the *same* auto-rater from one group to another, as a measure of its specific degradation in the face of low-quality references.

Table 9 shows the results. Indeed, our five string- and embedding-based baselines are much more sensitive to poor reference quality than any other auto-rater, by a significant margin. Likewise, on the other hand, the auto-raters that improve their performance the most on languages with poor references comprise of six of the "LLM as a Judge" models and four official submissions to the shared task: (MR7.2.1 and MR6 are based on Gemma 3 models; Q_MQM and Q_RELATIVE-MQM are based on Qwen 3.) All are reference-free systems, as expected.

4.4.5 Metric Score Difference Interpretation

Following the WMT Metrics Shared Task in the last two years, we continue to conduct analyses to find the threshold of metrics' score differences that corresponds to statistical significance of MT system rankings demonstrated by human annotators and the metrics themselves.[11] These analyses provide an interpretation of the metrics' score differences, support building an intuitive sense of metric score meanings, and encourage broader adoption of new automatic MT evaluation metrics. This year, since we expanded the number of language pairs (LPs) from 3 to 16, instead of analyzing metrics score differences by individual LP we are pooling the 14 LPs with references together in the following analyses for clear presentation and ease of understanding.

As a reminder, the results in this section should *not* be used as arguments to forego significance tests or appropriate human evaluation.

Correspondence to human scores significance: We first study the relationship between statistically significant differences in human scores and the magnitude of metric differences as in (Lo et al., 2023a). We run a one-sided paired t-test with an equal variance assumption for each system pair on segment-level human scores. After that, we fit the corresponding metric score differences and the p-values of the t-test on the human scores to an isotonic regression (Robertson et al., 1988), which predicts whether the human score difference will be significant given the metric's score difference. This year, we also consider the sign of the metric's difference. If the metric's decision disagrees with the human's but the human score difference is insignificant, we also consider that as a correct prediction. Isotonic regression produces a non-decreasing function where the classifier output can be interpreted as a confidence level.[12] We set $p_h < 0.05$ as the significance level of human scores. Thus, the output of the isotonic regression function can be viewed as $Pr(p_h < 0.05 \mid \Delta m)$ where p_h is the p-value of the t-test on the human scores for each system pair and Δm is the metric score difference.

Figure 2 shows the (log) p-value of one-sided paired t-test on the human scores against the corresponding BLEU, YISI-1, and TASER-REF score difference for each system pair. Additional figures (Figures 9-11 in Appendix C) show the same analyses for all metrics. For each metric, we can choose a particular level of confidence (i.e., a point along the y-axis on the right) to get the metric score difference cutoffs (i.e. a point along the x-axis) that this metric difference reflects significant human score differences. Drawing a horizontal line from the confidence level, say 80%, to the red line enables us to find the minimum metric difference cutoff required at the corresponding x-value down from the red line, i.e. 3.0 for BLEU in Figure 2. Using this lookup method, Table 10 show the cutoffs of Δm when $Pr(p_h < 0.05 \mid \Delta m) = 0.8$ for each metric. We run 10-fold cross-validation, and Table 10 shows that the range of precision in the cross-validation is consistently high across metrics. This means the metric cufoffs we find using the regression model are reliable.

Table 10 serves as a reference for understanding the score differences between MT systems provided by modern metrics. For example, we see that a BLEU difference of 3.0 corresponds to 80% confidence that two MT systems ranked by BLEU will match the decision made by human annotators with a significant difference. Meanwhile, a

[11] This section uses the term "metric," but the analysis is extended to auto-raters of all types as defined earlier in this paper.

[12] scikit-learn.org/stable/modules/isotonic.html

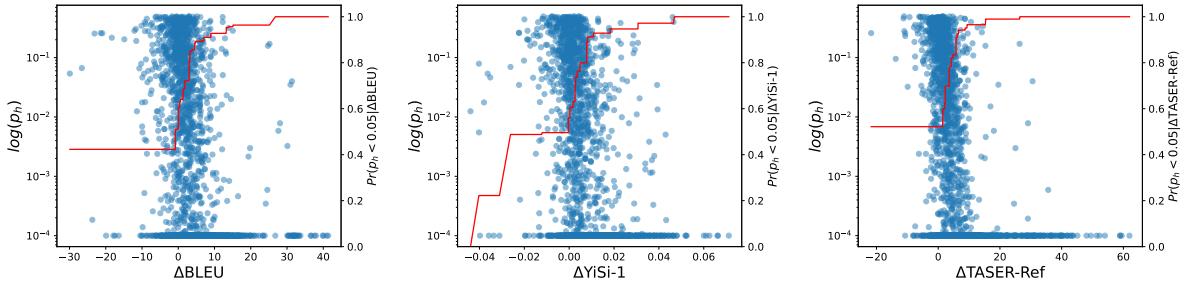


Figure 2: Log p-value of one-sided paired t-test on human scores (p_h) against each metric (left: BLEU, center: YISI-1, right: TASER-REF) score difference for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \,|\, \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001.

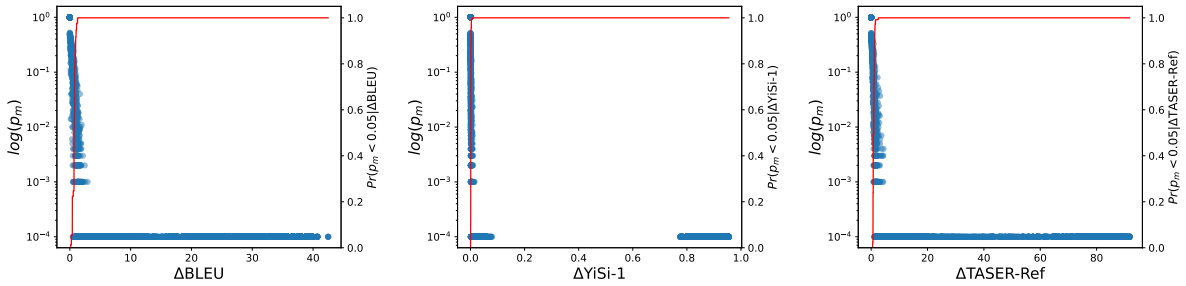


Figure 3: Log p-value of significance test with bootstrap resampling (p_m) on system-level metric scores against each metric (left: BLEU, center: YISI-1, right: TASER-REF) score difference for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \,|\, \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001.

TASER-REF score difference of 4.2 would have the same 80% chance of human-judged significant difference.

Correspondence to metric scores significance: We run a study similar to that above, but on the relations between statistically significant differences in metric scores and the magnitude of metric differences as inspired by Marie (2022). Instead of the one-sided t-test on human scores, the p-values are now obtained by running statistical significance tests with bootstrap resampling on the metric scores for each system pair. We fit the corresponding metric score differences and the p-values of the significance test to an isotonic regression for predicting whether the translation quality improvement as indicated by the metric will be significant given the metric score difference. We set $p_m < 0.05$, and thus the output of the isotonic regression function is now $Pr(p_m < 0.05 \,|\, \Delta m)$, where p_m is the p-value of the significance test on the metric scores for each system pair and Δm is the metric score difference.

Figure 3 shows the (log) p-value of the significance test with bootstrap resampling on the metric scores for BLEU, YISI-1 and TASER-REF score difference of each system pair. Additional figures (Figures 12-14 in Appendix C) show the same analyses for all metrics. Using the same lookup method described in the previous study, Table 11 shows the cut-offs of Δm when $Pr(p_m < 0.05 \,|\, \Delta m) = 0.8$ for each metric. We run 10-fold cross-validation, and Table 11 shows that the range of precision in the cross-validation is consistently high across metrics. This means that the metric cutoffs we find using the regression model are reliable.

Table 11 serves as a reference of metric differences that correspond to statistical significance with high confidence. For example, we see that a BLEU difference of 0.87 corresponds to 80% confidence that the difference is statistically significant. Meanwhile, a TASER-REF score difference of 1.1 would have the same 80% chance of being statistically significant. Our results, agreeing with Marie (2022), show that to claim significant differences ($p_m < 0.05$) on BLEU with high confidence (80%), the differences should be higher than the shared understanding (0.5 BLEU) in the research community.

We have to emphasize again that this result

Metric	minΔm	c.v. precision
Baselines		
YiSi-1	0.0078	[86-94%]
chrF	2.8	[89-93%]
spBLEU	4.0	[88-94%]
BERTScore	0.014	[87-94%]
BLEU	3.0	[86-92%]
COMET22	0.017	[84-92%]
sentinel-cand	0.23	[79-91%]
COMETKiwi22	0.048	[82-100%]
sentinel-src	—	—
Primary		
GEMBA-v2	2.1	[89-94%]
TASER-No-Ref	5.1	[93-97%]
rankedCOMET	0.057	[80-90%]
MetricX-25	2.9	[85-93%]
mr7_2_1	2.6	[86-93%]
SEGALE-QE	4.0	[83-97%]
Polycand-2	2.9	[82-93%]
Q_Relative-MQM	7.4	[87-94%]
EnsembleSlick	0.070	[55-100%]
hw-tsc	0.051	[83-100%]
UvA-MT	0.53	[74-100%]
Secondary		
TASER-Ref	4.2	[90-96%]
MetricX-25-Ref	2.3	[84-93%]
baseCOMET	0.017	[84-92%]
MetricX-25-QE	2.4	[84-91%]
mr6	2.2	[85-94%]
Q_MQM	1.9	[85-93%]
Polyic-3	3.2	[83-95%]
AutoLQA	0.015	[79-91%]
Polycand-1	2.6	[77-94%]
CollabPlus	0.025	[71-90%]
CollabSlick	0.043	[72-94%]
hw-tsc-max	0.061	[87-100%]
hw-tsc-base	0.052	[79-100%]
LLM-as-a-judge		
GPT-4_1	6.1	[89-95%]
CommandA	2.8	[86-93%]
Claude-4	3.7	[88-96%]
DeepSeek-V3	1.9	[86-95%]
Qwen3-235B	3.8	[89-95%]
Qwen2_5-7B	1.2	[82-91%]
AyaExpanse-32B	1.7	[83-95%]
Llama-3_1-8B	2.3	[75-95%]
Llama-4-Maverick	0.37	[79-91%]
CommandR7B	1.2	[76-92%]
Mistral-7B	3.5	[70-100%]
AyaExpanse-8B	0.77	[79-95%]

Table 10: Minimum Δm when $Pr(p_h < 0.05 \mid \Delta m) = 0.8$ for each metric in all language pairs with references (rounded to 2 significant figures), and the range of precision for the isotonic regression model in 10-fold cross-validation.

Metric	minΔm	c.v. precision
Baselines		
YiSi-1	0.0013	[98-100%]
chrF	0.66	[99-100%]
spBLEU	0.75	[99-100%]
BERTScore	0.0029	[99-100%]
BLEU	0.87	[99-100%]
COMET22	0.0041	[99-100%]
sentinel-cand	0.039	[99-100%]
COMETKiwi22	0.0046	[99-100%]
sentinel-src	0.00	[100-100%]
Primary		
GEMBA-v2	0.71	[99-100%]
TASER-No-Ref	1.2	[100-100%]
rankedCOMET	0.018	[100-100%]
MetricX-25	0.64	[99-100%]
mr7_2_1	0.82	[98-100%]
SEGALE-QE	0.95	[99-100%]
Polycand-2	0.47	[98-99%]
Q_Relative-MQM	2.1	[99-100%]
EnsembleSlick	0.0064	[99-100%]
hw-tsc	0.0060	[99-100%]
UvA-MT	0.030	[99-100%]
Secondary		
TASER-Ref	1.1	[99-100%]
MetricX-25-Ref	0.52	[99-100%]
baseCOMET	0.0041	[99-100%]
MetricX-25-QE	0.45	[99-100%]
mr6	0.85	[99-100%]
Q_MQM	0.53	[99-100%]
Polyic-3	0.43	[98-100%]
AutoLQA	0.0099	[99-100%]
Polycand-1	0.32	[98-100%]
CollabPlus	0.0079	[98-100%]
CollabSlick	0.0066	[99-100%]
hw-tsc-max	0.0056	[99-100%]
hw-tsc-base	0.0057	[99-100%]
LLM-as-a-judge		
GPT-4_1	1.5	[99-100%]
CommandA	0.85	[99-100%]
Claude-4	1.1	[99-100%]
DeepSeek-V3	0.61	[98-100%]
Qwen3-235B	0.87	[99-100%]
Qwen2_5-7B	0.85	[98-100%]
AyaExpanse-32B	0.55	[98-100%]
Llama-3_1-8B	1.6	[99-100%]
Llama-4-Maverick	0.74	[99-100%]
CommandR7B	0.90	[99-100%]
Mistral-7B	0.94	[98-100%]
AyaExpanse-8B	0.45	[98-100%]

Table 11: Minimum Δm when $Pr(p_m < 0.05 | \Delta m) = 0.8$ for each metric in all language pairs with references (rounded to 2 significant figures), and the range of precision for the isotonic regression model in 10-fold cross-validation.

should *not* be interpreted as evidence to forego significance test or appropriate human evaluation. Instead, we are only providing assistance to build an intuition on the meaning of the scores provided by the new metrics to encourage the transition away from lexical metrics towards more recent and stronger metrics.

5 Task 2: Span-Level Error Detection

This section presents the span-level error detection task in more detail. We discuss in more depth the error annotations per language pair (Section 5.1), and describe the baselines (Section 5.2) and the participant submissions (Section 5.3). Our meta-evaluation is described in Section 5.4. We then present the results, mostly focusing on the primary submissions, in Section 5.5.

5.1 Error Annotations

We use the ESA and MQM annotations sourced from the General MT task for this task as well, considering only the subset of documents and systems that were human-evaluated. We note that the error span patterns vary significantly per language as shown in Figure 4, which is a complementary view of Figure 1. For the translation pairs referring to lower-resource languages (e.g. English-Maasai), we frequently have the phenomenon where the whole segment is annotated as an error (frequently corresponding to hallucinated text). In contrast, annotations for higher-resource language pairs (e.g. Czech-German, English-Italian) correspond mostly to smaller, isolated error spans.

5.2 Baselines

XCOMET (Guerreiro et al., 2024) XL (3.5B) and XXL (10.7B) are neural models that are trained to identify MQM error spans in sentences along with a final quality score, thus leading to an explainable neural auto-rater. It adopts a unified input and output approach, allowing the prediction of translation quality assessment in multiple input modes (SRC-ONLY, SRC+REF and REF-ONLY), as well as generates sentence-level and word-level quality assessments. We use the SRC+REF mode with word-level predictions as the official shared task baselines.

Human2 For a subset of languages (except JA-ZH_CN and EN-KO_KR), the General MT shared task also collected a second round of human annotations. While not strictly a baseline, comparing submissions against a HUMAN2 set of evaluations provides additional insights into how automated metrics perform relative to human judgment. We report some statistics on HUMAN2 against HUMAN1 annotations in Table 12.

	# Errors	# Major Errors
Human1	33.56%	18.54%
Human2	32.48%	16.95%

Table 12: Translation error distribution on human annotations.

5.3 Submissions

We note that, this year, all task participants employed an LLM-based auto-rater to produce the fine-grained annotations. Specifically, the following systems were submitted to the task:

AIP (Yeom et al., 2025) The participants propose a tagged span annotation (TSA) approach, i.e., using reasoning LLMs to introduce inline numbered tags (e.g. $< v_0 > error\_span < \backslash v_0 >$) that explicitly mark error spans, and can easily map to diverse annotations (error severity, type, etc.) within the translated text. They enhance the tag schema to allow for annotation of omissions using zero-length tags. To be able to insert such tags on the hypothesis segments, they employed the OpenAI o3 and o4-mini reasoning models, leveraging the structured-output response mode to detect translation errors at the span level, formatted as JSON strings with the TSA approach mentioned above. They use few-shot examples and optimize for precision and minimality, explicitly prompting the models to (i) only label spans that it is confident are erroneous, and (ii) restrict the annotation to the minimal substring responsible for the error.

AutoLQA (Hrabal et al., 2025) The participants leveraged their Automatic Linguistic Quality Assessment (AutoLQA) systems, i.e., LLM-based evaluators designed to produce complete MQM-style annotations, including error spans, categories, and severities. The team fine-tuned GPT-4.1 and GPT-4o-mini model variants using internal data (≈100,000 segments), and using the WMT-QE-22 and Google-MQM datasets (dev + test) to determine performance improvements. They experiment with different prompts, controlling for the annotation structure, i.e., relaxing the MQM annotations to remove the category and approximate the ESA style. Their primary submission corresponds to the fine-tuned GPT-4o-mini and the secondary to the GPT-4.1-mini version, respectively.

GemSpanEval (Juraska et al., 2025) The participants fine-tuned a Gemma 3 27B model on past WMT MQM annotations formatted as JSON. Train-

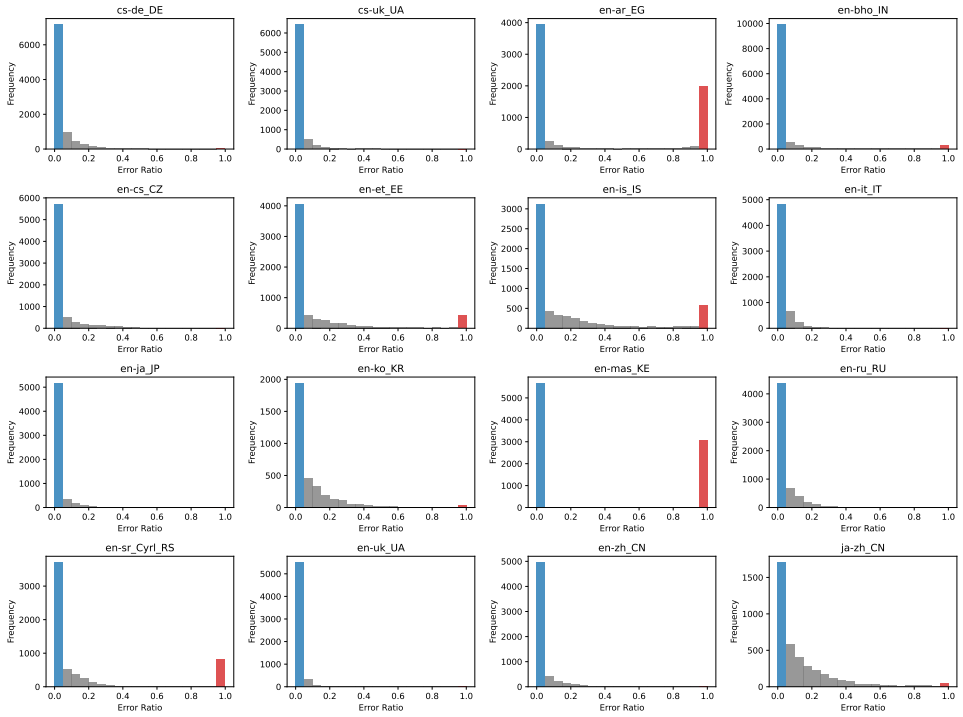


Figure 4: Distribution of error span ratio over the full segment length per language pair.

Language Pair	Baselines						Primary Submissions									Human2		
	XCOMET-XL			XCOMET-XXL			AutoLQA			AIP			GemSpanEval					
	P	R	f1	P	R	f1	P	R	f1	P	R	f1	P	R	f1	P	R	f1
CS-DE_DE	24.55	5.15	8.52	25.17	5.22	8.65	17.71	4.02	6.56	11.94	20.37	15.06	28.89	6.02	9.96	30.46	41.08	34.98
CS-UK_UA	25.02	4.00	6.90	28.21	3.56	6.32	20.73	1.93	3.54	13.60	10.16	11.63	35.94	3.58	6.52	27.67	28.95	28.30
EN-AR_EG	11.57	19.53	14.54	8.48	17.92	11.51	11.45	22.95	15.28	2.51	30.41	4.63	19.23	22.18	20.60	79.61	76.37	77.96
EN-BHO_IN	22.87	3.47	6.02	33.46	3.40	6.17	15.55	3.48	5.69	9.38	8.19	8.74	28.40	3.68	6.52	61.31	54.03	57.44
EN-CS_CZ	16.06	6.02	8.76	22.67	6.87	10.55	14.22	4.12	6.39	7.27	15.30	9.85	24.20	6.38	10.10	14.40	24.86	18.24
EN-ET_EE	14.93	16.66	15.75	16.56	17.07	16.81	14.84	11.82	13.16	5.71	20.34	8.92	23.09	12.28	16.04	33.31	32.87	33.09
EN-IS_IS	22.41	27.72	24.78	29.10	28.30	28.70	8.85	19.68	12.21	9.15	35.69	14.57	25.90	19.58	22.30	36.44	40.10	38.18
EN-IT_IT	30.60	4.16	7.33	23.40	5.40	8.77	17.89	2.55	4.47	10.45	13.70	11.86	33.71	5.47	9.41	30.52	30.62	30.57
EN-JA_JP	13.97	3.67	5.81	14.85	3.69	5.92	22.70	2.30	4.18	8.88	11.72	10.10	28.47	3.32	5.94	10.61	13.93	12.04
EN-KO_KR	8.74	14.64	10.95	9.96	17.09	12.58	20.23	7.26	10.69	4.81	25.89	8.12	17.65	10.54	13.20	-	-	-
EN-MAS_KE	10.23	34.84	15.81	11.35	36.31	17.29	15.14	38.95	21.80	27.94	28.65	28.29	35.03	35.67	35.35	94.73	92.14	93.41
EN-RU_RU	16.70	8.59	11.34	17.95	8.48	11.52	13.77	3.84	6.01	8.74	16.77	11.49	28.28	6.49	10.55	25.28	27.56	26.37
EN-SR_CYRL	21.18	21.59	21.38	24.17	21.07	22.52	13.61	15.16	14.35	6.81	27.11	10.88	21.66	15.67	18.18	61.67	58.32	59.95
EN-UK_UA	21.84	2.98	5.25	27.31	3.20	5.73	15.48	1.32	2.43	12.63	6.98	8.99	37.01	2.19	4.13	34.76	39.28	36.88
EN-ZH_CN	22.62	3.80	6.50	19.45	4.14	6.83	13.08	2.92	4.78	7.72	10.43	8.87	30.02	3.37	6.07	11.84	12.82	12.31
JA-ZH_CN	26.89	21.80	24.08	24.07	20.04	21.87	14.83	13.58	14.18	8.47	41.64	14.08	25.35	17.46	20.68	-	-	-
Average	19.39	12.41	12.11	21.01	12.61	12.61	15.63	9.74	9.11	9.75	20.21	11.63	27.68	10.87	13.47	47.04†	48.31†	47.48†

Table 13: Task 2 micro-F1 (%) by language pair for all auto-raters. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

ing data covered the period WMT20–24 (Specia et al., 2020, 2021; Zerva et al., 2024), and optimization was performed with the Adafactor (Shazeer and Stern, 2018). A central focus of their approach was the resolution of error span ambiguity, ensuring that predicted spans are uniquely identifiable within the hypothesis segment. The model was trained to extend spans with additional context whenever a substring was not unique. The context expansion covers both preceding and following context and proceeds incrementally — word by word for alphabetic languages and character by character for logographic or syllabic languages such as Chinese and Japanese — until a unique substring is obtained. The model was designed to operate in both reference-based and reference-free (QE) modes, and the team submitted both variants as their primary and secondary systems, respectively.

Language Pair	Baselines		Primary Submissions			Secondary Submissions			Human2
	XCOMET-XL	XCOMET-XXL	AutoLQA	AIP	GemSpanEval	AutoLQA-4.1	AIP	GemSpanEval-QE	
CS-DE_DE	13.07	16.22	13.91	36.45	17.08	16.63	31.22	19.14	64.46
CS-UK_UA	17.49	19.37	11.44	37.74	15.67	11.99	32.48	16.33	67.55
EN-AR_EG	10.54	10.07	11.53	18.86	12.24	10.12	14.04	12.89	79.33
EN-BHO_IN	20.00	9.88	7.14	22.44	7.01	5.46	13.40	7.71	78.86
EN-CS_CZ	10.79	10.93	17.36	31.78	12.68	13.14	25.72	14.28	60.70
EN-ET_EE	11.62	13.72	12.97	24.89	10.83	11.43	18.18	11.17	64.77
EN-IS_IS	15.87	18.50	10.01	21.83	14.95	9.59	17.22	15.24	62.51
EN-IT_IT	7.65	11.41	11.51	32.03	11.92	11.93	29.34	11.91	52.23
EN-JA_JP	10.67	16.15	11.50	44.81	8.33	10.39	37.91	12.15	64.29
EN-KO_KR	12.27	14.32	14.05	26.44	11.77	15.36	24.98	13.27	-
EN-MAS_KE	49.07	49.19	27.42	36.13	31.45	26.88	15.59	31.59	96.33
EN-RU_RU	13.19	13.80	19.09	30.50	10.77	18.06	27.29	11.20	58.49
EN-SR_CYRL	14.08	15.19	18.72	23.76	12.05	16.94	15.59	12.22	64.19
EN-UK_UA	10.17	10.79	16.97	33.69	6.55	10.07	27.20	6.37	72.02
EN-ZH_CN	6.55	8.65	32.37	38.14	7.50	32.29	33.65	8.01	59.82
JA-ZH_CN	20.17	19.78	18.90	25.83	25.74	18.48	25.03	26.76	-
Average	15.20	16.12	15.93	30.33	13.53	14.92	24.30	14.39	71.60†

Table 14: Task 2 macro-F1 (%) by language pair for all auto-rater submissions. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

5.4 Meta-Evaluation

For Task 2, we use the micro-F1 score between the predicted and the gold error spans calculated at the character level as the primary metric. The score is weighted to allow for half points for correctly identified spans with misclassified severity. Compared to the previous year, instead of computing the best matching annotation for each character (Zerva et al., 2024), we compute F1 over multiple error annotations per character, allowing for separate comparisons for each overlapping error span.

More specifically, for each hypothesis, we compute the counts for the number of "major" and "minor" errors at each character index separately for both gold annotations and predictions. This results in four statistics per hypothesis: gold major counts, gold minor counts, predicted major counts, and predicted minor counts, each of length equal to the length of the hypothesis. We then calculate a true positive (TP) score by iterating through each character position and assigning full credit based to the number of overlaps between gold and predicted counts of the same severity type (major with major, minor with minor) at each character. In the case of overlapping annotations with different severity, we assign a partial credit to the *unmatched* gold counts and predicted counts at the same character position, regardless of the original severity. This allows a predicted major error to get partial credit if it aligns with a character that was part of a gold minor error, and vice-versa. These TP scores are summed across all characters and all hypotheses. Finally, precision (P), recall (R), and F1 score are calculated based on the aggregated TP, total gold counts, and total predicted counts. The complete logic can be seen in Algorithm 1.

5.5 Main Results

Table 13 presents the complete results for all evaluated systems. Below are our primary observations and findings from these performance results.

Current auto-raters fail to localize errors. Across all language pairs where HUMAN2 scores are available, there is a very large gap between the auto-raters' performance scores and the human rater scores. HUMAN2 scores range from around 12% to over 93%, while the best auto-raters rarely exceed 35%, indicating the task is very challenging for current automated methods.

There is large variation across language pairs. No single auto-rater consistently outperforms others across all language pairs. The range of scores across different language pairs suggests varying levels of difficulty for the auto-raters. For example, most systems struggle significantly with EN-UK_UA, yielding very low F1 scores. In contrast, EN-MAS_KE allows the GemSpanEval system to achieve its peak scores. On average, GemSpanEval achieves the highest micro-F1 score (13.47%). AIP follows next with decent average performance (11.63% Primary). AutoLQA systems have the lowest average scores.

Auto-raters exhibit different precision-recall tradeoffs The AIP submissions, unlike all others, seem to be obtaining higher micro-recall, at the cost of lower micro-precision, despite including precision-focused instructions in the prompt.[13] This outcome contrasts sharply with XCOMET, AutoLQA/ESA, and GemSpanEval, which tend to be more conservative, often achieving higher precision than recall, especially on difficult languages. Human2, on the other hand, shows that high performance requires excelling in both measures, a balance that the auto-rater systems currently fail to achieve.

Auto-raters show different strengths in generalization and consistency. Table 14 reports macro-F1 score for all primary and secondary submissions. Similar to the primary evaluation (macro-F1), HUMAN2 achieves the best scores across the board. Interestingly, AIP stands out as the best submission in terms of macro-F1, averaging 30.33%. This is significantly higher than AutoLQA (15.93%) and GemSpanEval (13.53%). This largely suggests that AIP can achieve good F1 scores on average across language pairs, but its overall performance on the sheer volume of errors might be hampered by poor performance on certain heavily-weighted error segments or language pairs. For example, EN-AR_EG has many segments with full spans marked as errors due to hypotheses being in the wrong dialect (see Figure 4), and the gap in micro- and macro-F1 is large (14.23%). On the other hand, GemSpanEval's micro-F1 (13.47) is very close to its macro-F1 (13.53), which suggests a more consistent performance across language pairs in terms of the number of errors.

Overall, the results suggest that precisely locating error spans remain a challenging problem for auto-rater systems.

6 Task 3: Quality-Informed Segment-Level Error Correction

The subtask received a total of 6 submissions, from 3 participants. The results depict a clear outcome in terms of the winning system.

6.1 Data and Baselines

We reduced the number of language pairs to 6, and overall data size to a total of 6,000 instances

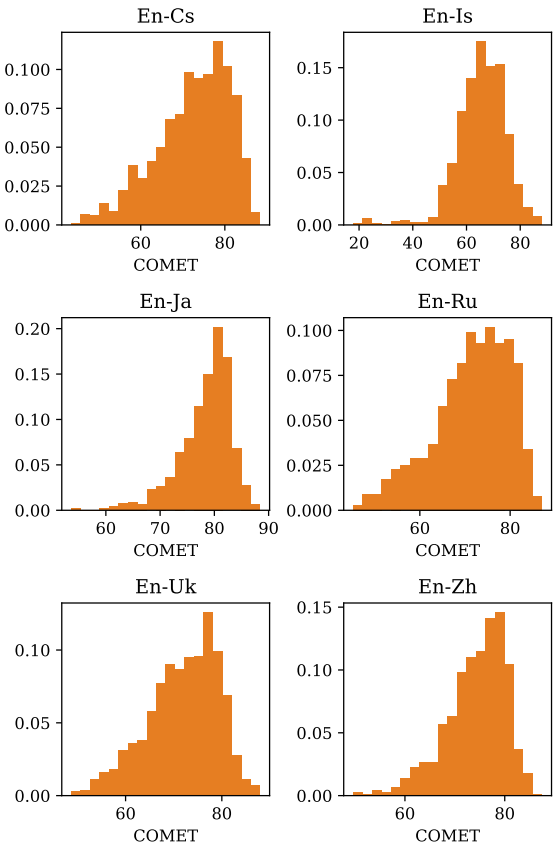


Figure 5: Task 3 - Language-pair-specific COMET score distribution of translations in the evaluation set

with an equal number of samples *per* pair. The preparation of the test set ensured representation of samples across all quartiles of the COMET score distribution. Due to Codabench limitations and memory-intensive COMET models used for evaluation, we displayed leaderboard results using a fixed 100 random samples from each submission during the competition phase. Figure 5 shows the COMET score distribution of the evaluation set of each language pair. We observe that for all language pairs, we have similar distributions skewed towards high-quality scores.

We include the following baselines along with the participants' submissions:

- **BASELINE-S1** This baseline translates the source segment from scratch using Gemma3-it-27B LLM (Gemma Team et al., 2025). It ignores the MT system output, and retranslates the input source segment. We provide the prompt used in Appendix Table 25.
- **BASELINE-S2** The final baseline uses QE information from post-edit original translation based on quality estimation from XCOMET-

[13] Table 23 shows that AIP achieves higher precision than recall on macro-F1 scores, which is aligned with their focus on being precision-focused at the instance level.

XL (Guerreiro et al., 2024), and then uses Gemma3-it-27B for APE. We provide the prompt used in Appendix Table 26.

6.2 Submissions

PHRASE (Hrabal et al., 2025) Participants leverage GPT models, specifically, o3 and o3-mini, to produce corrections over MT output given the source segment, without using the provided QE information. They use a proprietary common prompt for all systems submitted, and add variations to the prompt to change the correction strategy. The three proposed training-free approaches focus on either "only correcting errors (-S3)", "improving fluency (-S2)", and "improving fluency with steps to reason for corrections (-S1)".

SURREYPAI (Padmanabhan, 2025) Participant proposed two training-free approaches to Quality-informed error correction. The first approach (-S1) leverages the provided DA score, uses it as a selector, and routes to a specific open-weight LLM for re-translation using input QE information. This approach leverages one of six selected open-weight LLMs, wherein some LLMs were selected for their robust performance on other NLP tasks in the target language. The second approach (-S2) uses fine-grained error span information to replace an erroneous token with "__BLANK__" and then uses an LLM to replace this token contextually and "fills in the blank".

PACIFICO (Sharma, 2025) Participant proposed using natural language explanations as an intermediate step to the "detector-corrector" approach, which proposes error identification and then error correction. They use xTower to generate intermediate natural language explanations based on input QE information. The approach then feeds the explanation along with the source segment and MT output to the Gemini-1.5-Pro model to obtain final corrections.

6.3 Evaluation Metrics

We evaluate the quality of the corrections over MT, using ΔCOMET as the primary metric, and Gain-to-Edit Ratio (GER) to quantify efficiency.

ΔCOMET: Measures how much the in-post-edits improve over the original MT output (hyp) based on COMET score (Rei et al., 2022b). COMET[14] is a neural evaluation metric trained on human quality assessments, designed to capture meaning preservation and fluency by comparing translations against the source:

$$\Delta\text{COMET} = \text{COMET}(\text{src}, \text{pe}) - \text{COMET}(\text{src}, \text{hyp})$$

Positive values signal that post-editing yields a translation judged closer to human quality, while negative values imply a degradation relative to the initial MT output.

Gain-to-Edit Ratio: This metric evaluates the efficiency of edits by relating quality gains to the editing effort. According to our formulation, it is defined as the ratio between ΔCOMET and the Translation Edit Rate (TER)[15] (Snover et al., 2006) between the post-edited output (pe) and the original MT output (hyp):

$$\text{Gain-to-Edit Ratio} = \frac{\Delta\text{COMET}}{\text{TER}(\text{pe}, \text{hyp})}$$

Higher values indicate that larger quality improvements are achieved with fewer edits, while lower or negative values suggest limited or detrimental improvements relative to the editing cost.

6.4 Main Results

The main results for Task 3 are summarized in Table 15, and other metrics used for analysis are reported in Table 16. Submissions are ranked primarily by the average ΔCOMET across languages (Table 15). SURREYPAI-S1 attains the best system-wide performance, leading on both ΔCOMET for every language pair; PHRASE-S1 stands at the next best, and these two are the only submissions that surpass the BASELINE-S2 results over the primary metrics. However, in terms of efficiency of edits (GER), PHRASE-S1 obtains a higher score for En-Is, and BASELINE-S2 seems to perform the best for En-Uk.

Figure 6 illustrates the mean change in ΔCOMET for eight different systems, including two baselines, across six target languages. A clear finding is the superior performance of the SURREYPAI-S1 system, which consistently achieves a positive ΔCOMET across all language pairs, indicating an improvement in translation quality. This system shows particularly strong gains for Icelandic (`is_IS`) and Russian (`ru_RU`).

[14] Unbabel/wmt22-cometkiwi-da

[15] Computed using TERCOM-0.7.25 with default flags except the case sensitivity. Character-level tokenization is used for Chinese and Japanese, and *sacrebleu* (Post, 2018) '13a' for the rest.

System Name	En-Cs		En-Is		En-Ja		En-Ru		En-Uk		En-Zh		Average	
	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER	ΔCOMET	GER
SURREYPAI-S1	**0.019**	**0.015**	**0.037**	0.027	**0.010**	**0.008**	**0.020**	**0.016**	**0.016**	**0.012**	**0.018**	**0.015**	**0.020**	**0.016**
PHRASE-S1	0.003	0.006	0.032	0.058	-0.006	-0.012	-0.004	-0.007	-0.003	-0.006	-0.002	-0.005	0.003	0.006
BASELINE-S2	0.000	0.000	0.007	0.026	-0.008	-0.036	0.002	0.009	0.004	0.017	-0.005	-0.023	0.000	-0.001
BASELINE-S1	-0.002	-0.002	0.008	0.005	-0.005	-0.004	-0.001	-0.001	-0.003	-0.002	0.002	0.002	0.000	0.000
PACIFICO	-0.008	-0.032	0.019	0.054	-0.018	-0.085	-0.016	-0.061	-0.008	-0.034	-0.007	-0.036	-0.006	-0.033
PHRASE-S3	-0.008	-0.030	0.027	**0.063**	-0.016	-0.060	-0.019	-0.056	-0.016	-0.045	-0.006	-0.023	-0.006	-0.025
PHRASE-S2	-0.011	-0.027	0.025	0.050	-0.018	-0.049	-0.024	-0.050	-0.020	-0.043	-0.009	-0.026	-0.010	-0.024
SURREYPAI-S2	-0.007	-0.005	-0.010	-0.006	-0.013	-0.008	-0.008	-0.005	-0.014	-0.009	-0.013	-0.010	-0.011	-0.007

Table 15: Task 3 - Performance of systems across languages with ΔCOMET and Gain to Edit Ratio (GER) metrics. Systems are ranked in order of average ΔCOMET.

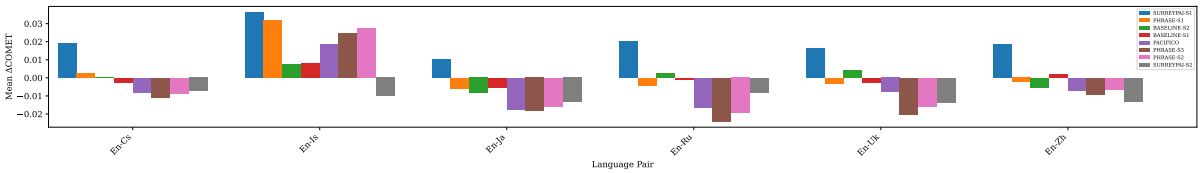


Figure 6: Task 3 - Mean ΔCOMET scores *per* **language pair** across submissions.

In contrast, most other systems exhibit mixed or negative results, with PHRASE-S2 systems frequently showing a degradation in quality, especially for Russian and Ukrainian (uk_UA). Interestingly, the English–Icelandic language pair exhibits the most notable overall improvements, with several systems achieving consistent gains. In contrast, performance on English–Japanese remains relatively limited across all systems. Correlating these outcomes with the corresponding evaluation sets suggests that LLM-based approaches follow trends observed in earlier transformer encoder–decoder APE systems (Akhbardeh et al., 2021; Bhattacharyya et al., 2023; Zerva et al., 2024). Specifically, when baseline translations are weaker—evidenced by TER distributions skewed toward higher values, APE systems tend to yield larger improvements, and vice versa. A similar pattern is observed with COMET: When the score distribution is skewed toward the upper end (Figure 5), the marginal improvements achievable by LLM-based automatic post-editing systems tend to diminish.

Figure 7 indicates that while SURREYPAI-S1 improves translations across all four domains, PHRASE-S1 shows mildly positive or negligible improvements on all. Interestingly, PACIFICO shows decent gains on *literary* and *social* domain data, but degradation in performance on the *news* and *speech* data, leads to its lower rank on the overall results. It also indicates that improvements in the *speech* domain are tough to obtain and no other system, except SURREYPAI-S1, shows improvements in terms of translation quality. We note that the speech domain data is derived from ASR transcriptions, indicating that text data derived from multimodal input may need further investigation or a different approach to correction.

Edit-Operations Figure 8 illustrates the distribution of post-editing operations like insertion (green), deletion (blue), substitution (orange), and shift (red) across various systems for six different language pairs. A clear and consistent trend is observable across all conditions: *substitution* is the most frequent edit operation, typically accounting for more than 50% of all changes. This suggests that the primary challenge for the translation systems lies in lexical choice rather than fluency. Deletion is generally the second most common error, followed by insertion. Shift operations, which correct word order, are consistently the least frequent type of edit, indicating that the models generally produce syntactically plausible translations. While this distribution pattern holds for all systems and language pairs, there are subtle variations; for instance, translations into typologically distant languages like Japanese and Chinese appear to necessitate a mildly higher proportion of insertions and deletions compared to the other language pairs.

6.5 Meta-Evaluation Metrics

While the main evaluation relies on reference-less evaluation through ΔCOMET and GER, we complement them using reference-based metrics for further analysis. In particular, we adopt ΔBLEURT (Sellam et al., 2020), a neural metric that captures semantic similarity, and

Figure 7: Task 3 - Mean ΔCOMET scores *per* **domain** across submissions.

System Name	En-Cs		En-Is		En-Ja		En-Ru		En-Uk		En-Zh		Average	
	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF++	ΔBLEURT	ΔchrF/ΔchrF++
SURREYPAI-S1	-0.002	-3.460	0.053	0.000	0.003	0.000	0.009	0.000	0.002	0.000	-0.003	0.000	0.010	-0.577
PHRASE-S1	-0.012	2.256	0.031	2.856	-0.033	-0.558	-0.030	-11.636	0.023	13.758	-0.045	-0.279	-0.011	1.066
BASELINE-S2	-0.027	0.059	0.025	0.264	-0.007	-2.272	-0.006	0.733	-0.002	0.214	-0.007	-1.086	-0.004	-0.348
BASELINE-S1	-0.019	8.902	0.012	2.199	-0.009	4.358	0.011	0.954	0.018	7.267	0.007	-1.662	0.003	3.670
PACIFICO	-0.035	8.546	-0.004	10.434	-0.030	2.934	-0.039	-42.936	-0.024	6.791	-0.010	5.989	-0.024	-1.374
PHRASE-S3	-0.110	6.550	-0.075	7.117	-0.053	-0.579	-0.210	-14.537	-0.171	8.471	-0.090	0.331	-0.118	1.226
PHRASE-S2	-0.104	4.810	-0.076	5.120	-0.050	-2.228	-0.206	-16.787	-0.176	3.223	-0.083	4.760	-0.116	-0.184
SURREYPAI-S2	-0.015	-1.219	0.002	-5.248	-0.024	0.029	-0.007	0.041	0.002	-22.708	-0.024	0.000	-0.011	-4.851

Table 16: Task 3 - Performance of participant systems across languages with ΔchrF for En-Ja, En-Zh, ΔchrF++ for the rest, and ΔBLEURT for all language pairs.

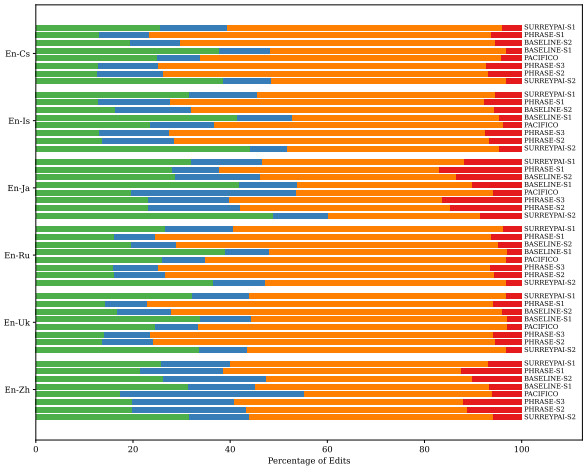


Figure 8: Task 3 - Language pair-wise **distributions of edit operations** performed on the original translation. Green indicates insertion, Blue indicates deletion, Orange indicates substitution, and Red indicates shift operations.

ΔchrF++[16] (Popović, 2017), a character n-gram overlap metric that emphasises lexical similarity. Unlike the primary reference-less metrics, which measure how much system outputs diverge from the raw translations in the desired direction, these allow us to have *an indication of how much closer system outputs have moved toward the reference or gold-standard translations* in terms of semantics and lexical distance.

Δ BLEURT: BLEURT is a learned, reference-based evaluation metric that leverages pretrained language models fine-tuned on human-rated data to capture semantic adequacy and fluency beyond surface overlap (Sellam et al., 2020). According to our formulation, ΔBLEURT measures the change in BLEURT when moving from the original MT output (hyp) to the post-edited output (pe) against the same reference (ref):

$$\Delta\text{BLEURT} = \text{BLEURT}(\text{pe, ref}) - \text{BLEURT}(\text{hyp, ref})$$

A positive ΔBLEURT indicates that resultant corrections improve semantic similarity to the reference, while a negative value suggests a reduction in quality.

ΔchrF++: chrF++ computes an F-score over word-level n-gram precision and recall between the hypothesis translation and a reference (Popović, 2017). The ΔchrF++ captures word n-gram quality gains, providing an additional cross-check beyond semantic metrics such as COMET.

$$\Delta\text{chrF++} = \quad \text{chrF++}\,(\text{pe, ref}) - \text{chrF++}\,(\text{hyp, ref})$$

A positive ΔchrF++ indicates that corrections improve similarity to the reference, while a negative value suggests a reduction in quality. We report ΔchrF for Chinese and Japanese and ΔchrF++ for the rest.

6.6 Meta-Evaluation Results

From Table 16, we observe that gains are mixed across participant systems and languages. Overall, ΔBLEURT improvements are only visible for SURREYPAI-S1 and BASELINE-S1, indicating that corrections did not substantially improve semantic adequacy across most systems. Multiple systems perform strongly on the English–Icelandic pair, particularly in ΔchrF++,

[16]Computed using sacrebleu (Post, 2018) with default flags

while English–Chinese exhibits degradation for nearly all submissions, and English–Czech results remain mixed. On English–Ukrainian, PHRASE-S1 and BASELINE-S1 clearly outperform SURREYPAI-S1 by a wide margin.

BASELINE-S1 achieves the highest system average for ΔchrF++, with PHRASE-S3 and PHRASE-S1 also showing moderate positive gains. In contrast, no other systems achieve consistent improvements in overall translation quality. Among baselines, BASELINE-S1 is comparatively closer to references, delivering competitive results across multiple languages, whereas BASELINE-S2 remains close to neutral.

Several systems (PACIFICO, PHRASE-S2, SURREYPAI-S2) show negative deltas on the secondary metrics for some languages, suggesting that their edits often diverge from reference translations despite attempted corrections. Notably, PACIFICO displays extreme variance—achieving large gains on English–Icelandic and English–Chinese, but severe degradation on English–Russian (–42.936 ΔchrF++), indicating the instability of certain LLM-based approaches across language pairs.

A closer comparison of both SurreyPAI submissions shows contrasting behaviour: SURREYPAI-S1 achieves the best system average in ΔBLEURT (+0.010), suggesting modest improvements in semantic adequacy, but its chrF average remains negative. In contrast, SURREYPAI-S2 underperforms on both metrics, with the steepest degradation in chrF (–4.851), particularly on the English–Ukrainian pair (–22.708), highlighting the sensitivity of system design choices to specific language pairs. Given the approach to SURREYPAI-S1, the system is able to retranslate and improve on translation quality, but Table 16 shows that such an approach does not bring the output closer to a known reference. At a language level, English-Icelandic seems to show the most consistent improvements, while English-Russian shows the most severe degradations. English-Chinese and English-Japanese also prove challenging, with limited gains, which may suggest that languages with morphological richness (Ukrainian and Icelandic) may offer opportunities for effective corrections, whereas typologically distant languages (Chinese and Japanese) are still harder to handle.

We also conducted batchwise significance testing with ΔCOMET scores to compare system performance. The dataset with $6,000$ instances was divided into 60 fixed batches with unique samples and 100 additional randomly sampled batches, yielding a total of 160 batchwise averages per system. For each batch, the system with the highest ΔCOMET score was identified, and the frequency of these "wins" across all batches served as an indicator of each system's consistency and robustness. Table 17 shows SURREYPAI-S1 dominates the evaluation with 157 out of 160 wins, while all other systems achieved at most two wins (PHRASE-S2 with 2 and PHRASE-S1 with 1), and the remaining systems failed to win a single batch. Testing reveals SURREYPAI-S1 to be the best system consistently across both static and random batch settings.

System	Win Count
SURREYPAI-S1	157
PHRASE-S2	2
BASELINE-S2	0
BASELINE-S1	0
PHRASE-S1	1
PACIFICO	0
PHRASE-S3	0

Table 17: Task 3 - Batch-wise meta-evaluation: winning counts per system.

6.7 Task Overview

This sub-task marks the first instance where translations were majorly generated by LLMs rather than by traditional NMT systems. We observe that LLM-based APE systems struggle to further improve these translations. This trend is analogous to earlier iterations: while neural APE systems could successfully enhance SMT outputs, they initially faced difficulties in improving NMT-generated translations. Then last year, LLM-based APE systems demonstrated the ability to improve NMT translations even for underrepresented languages. In contrast, when confronted with LLM-generated translations, even in high-resource languages, they now appear to encounter similar challenges. It requires innovative and sophisticated strategies that can effectively address the unique challenges inherent in the high-quality translations produced by LLMs.

7 Challenge Sets

For the third year, our shared task included a subtask involving challenge sets. This subtask is inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017), which aimed at testing the generalizability of NLP systems beyond the distributions of their training data.

Challenge Set	LPs	Phenomena	Items
CoDrift (Tan et al., 2025)	3	continuation drift	3,326
GAMBIT+ (Filandrianos et al., 2025)	33	gender bias	289,443
MSLC25 (Knowles et al., 2025)	2	low quality MT	369
SSA-MTE (Li et al., 2025a)	11	African languages	12,769

Table 18: Overview of the participation at the metrics challenge sets subtask.

Whereas the standard evaluation of the shared task is conducted on test sets containing generic text from real-world content, the challenge set evaluation is based on test sets designed with the aim of revealing the abilities or the weaknesses of the metrics or evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the subtask takes place in a decentralized manner: contrary to the main metric task, the test sets are not provided by the organizers but by different research teams, who are also responsible for analyzing and presenting the results.

7.1 Subtask Structure

This subtask is made of three consecutive phases; (1) the *Breaking Round*, (2) the *Scoring Round*, and (3) the *Analysis Round*:

1. In the *Breaking Round*, every challenge set participant (*Breaker*) submits their challenge set S composed of examples for different phenomena, where every example $(s, t, r) \in S$ contains one source sentence s, one translation hypothesis t, and one reference r.

2. In the *Scoring Round*, the metrics participants from the main task (the *Builders*) are asked to score with their metrics the translations in the given test set. Also, in this phase, the metrics task organizers score all data with the baseline metrics.

3. Finally, after having gathered all metric scores, the organizers return the respective scored translations to the *Breakers* for the *Analysis Round*, where they employ their own evaluation for the performance of the metrics with regard to the phenomena they intended to test.

7.2 Challenge Set Descriptions

This year there were 4 submissions, covering a wide range of phenomena and 23 different language pairs, which supersede the official language pairs of the Metrics Shared Task. An overview of the submitted challenge sets can be seen in Table 18. A short description of every submission follows:

CoDrift (Tan et al., 2025) Quality Estimation (QE) models such as COMET-KIWI, MetricX, and ReMedy exhibit a recurring failure mode: they often assign high scores to translations that start faithfully but subsequently drift into fluent yet irrelevant content. To systematically investigate this issue, Tan et al. (2025) present CoDrift, a WMT25 challenge set designed to stress-test QE robustness against continuation drift. The dataset is constructed entirely from controlled large language model (LLM) experiments: for each source sentence, we generate multiple "drift" candidates whose continuation length and topical proximity are systematically manipulated. This design enables precise control over the degree of semantic divergence, while maintaining surface fluency, thereby creating challenging cases that can mislead current QE systems. CoDrift aims to provide the community with a targeted benchmark for diagnosing and improving QE models in the presence of subtle off-target content.

Gambit+ (Filandrianos et al., 2025) In this submission, the authors introduce GAMBIT+[17], a large-scale challenge set designed to probe gender bias in QE systems. The dataset extends the GAMBIT corpus of English gender-ambiguous occupational terms to three source languages (English, Turkish, Finnish), where occupational gender is not specified, and 11 target languages with grammatical gender: Arabic, Czech, Greek, Spanish, French, Icelandic, Italian, Portuguese, Russian, Serbian, and Ukrainian. Importantly, all occupations are linked to the ISCO-08 classification, an internationally recognized standard for categorizing jobs, which enables fine-grained per-occupation analysis and ensures coverage of the full occupational spectrum. For each source text, two parallel target translations were produced, one masculine and one feminine, differing only in the gender of the occupation and all dependent grammatical elements (e.g., pronouns, adjectives) to ensure consistency. An unbiased auto-rater should assign near-identical scores to both versions. Each source-target language pair contains over 8,500 source texts, with two parallel target translations (masculine and feminine),

[17] huggingface.co/datasets/ailsntua/gambit-plus

resulting in more than 17,000 source-translation pairs per language pair and over 550,000 pairs in total across the 33 language combinations. With its scale, full ISCO coverage, and strictly parallel design, GAMBIT+ provides a comprehensive and controlled resource for investigating gender fairness in QE metrics.

The authors benchmarked three baseline metrics and eight shared task submissions on GAMBIT+, though one baseline was excluded from the analysis since it evaluated only source texts rather than target translations. Across the remaining auto-rater systems, all showed statistically significant differences between masculine and feminine outputs, but the scale of these differences varied widely. For instance, UvA-MT and rankedCOMET displayed average normalized score gaps of over 100% and 70% respectively, while Polycand variants and Polyic metrics registered less than 4%. Bias magnitude was influenced by both the source and target languages, with English sources and target languages such as Arabic, Russian, and Icelandic exhibiting stronger disparities. At the occupational level, most auto-raters favored masculine translations overall, yet stereotypically female-associated roles (e.g., nursing, midwifery, cleaning professions) often saw the opposite pattern, reflecting known tendencies in MT systems. These results show that QE systems are sensitive to gender even in cases where they shouldn't be, amplifying occupational stereotypes rather than remaining neutral, underscoring the need for systematic auditing and fairness-aware design.

MSLC25 Challenge Set (Knowles et al., 2025) Based on the past two iterations of the Metric Score Landscape Challenge (MSLC; Lo et al., 2023b; Knowles et al., 2024), MSLC25 is a smaller-scale study of auto-rater performance on a broad range of MT quality along with several specific corner cases and phenomena. MSLC25 includes a collection of low- to medium-quality MT systems' output on Japanese–Chinese news data from the WMT25 General MT Shared Task test set. As in previous editions, the challenge set explores auto-rater scores assigned to empty strings in the source or target, showing unexpected results for some auto-rater systems. In small-scale proof-of-concept experiments (using Japanese, Chinese, English, and Czech data) the challenge set also examines auto-rater scores assigned to mixed- and wrong-language text and English language spelling variants. The results of MSLC25 continue to highlight the need for auto-rater builders to test their systems on corner cases and wide ranges of MT quality before releasing them to the broader research community.

SSA-MTE Challenge Set (Li et al., 2025b,a) The SSA-MTE challenge set is a large-scale benchmark for machine translation evaluation in Sub-Saharan African languages. It comprises 12,768 human-annotated adequacy scores across 11 language pairs involving English, French, and Portuguese, evaluated on outputs from six commercial and open-source machine translation systems. Results indicate that correlations with human judgments remain generally low, with most systems achieving Spearman correlations below the 0.4 threshold for medium-level agreement. Performance varies substantially across language pairs, and in extremely low-resource cases such as Portuguese–Emakhuwa, correlations drop to around 0.1, underscoring the challenge of evaluating MT for very low-resource African languages. Notably, the long-standing baseline metric chrF (Popović, 2015) achieves performance comparable to the strongest neural supervised submission, MetricX-25 (Juraska et al., 2025), an encoder-only regression model initialized from *Gemma3* (12B) (Gemma Team et al., 2025) and fine-tuned on WMT15–23 DA and MQM scores. However, these findings still highlight the urgent need for more robust and generalizable machine translation evaluation methods tailored to under-resourced African languages.

7.3 Challenge Set Results Overview

The studies collectively reveal critical weaknesses in current automatic MT evaluation systems. CoDrift shows that popular QE models like COMET-KIWI and MetricX often fail when translations drift into fluent but semantically irrelevant continuations, highlighting the need for robustness against subtle off-target content. GAMBIT+ uncovers systematic gender bias in QE systems across 33 language combinations, with some auto-raters showing over 100% score gaps between masculine and feminine translations, amplifying occupational stereotypes. MSLC25 emphasizes that auto-raters can behave unpredictably on low- to mid-quality outputs and corner cases such as empty strings or mixed-language outputs, stressing the importance of thorough auto-rater testing for real-world

robustness. Finally, SSA-MTE demonstrates that auto-rater correlations with human judgments remain very low for Sub-Saharan African languages, especially in extremely low-resource pairs, underscoring the urgent need for inclusive, generalizable evaluation methods.

8 Conclusion

This paper documented the results of the WMT25 shared task on automated machine translation evaluation systems, which unified the Metrics and QE Shared Tasks from previous years. The shared task this year consisted of three subtasks: (1) segment-level quality score prediction, (2) span-level translation error annotation, and (3) quality-informed segment-level error correction. Task 1 results indicate the strong performance of large LLM-as-a-judge auto-rater systems at the system level, while reference-based baseline metrics outperform LLMs at the segment level. Task 2 results indicate that accurate error detection and balancing precision and recall are persistent challenges. Task 3 results show that minimal editing is challenging even when informed by quality indicators. Robustness across the broad diversity of languages remains a major challenge across all three subtasks. As described throughout the paper, this year marked significant changes to multiple dimensions of the evaluation. Evaluation data, originating from the General-MT task, was more challenging for MT systems, and covered a diverse set of new language-pairs. The move to long segments and the adoption of ESA human annotation for most of the languages were also new. We strongly believe that these changes were all warranted by the changing landscape in the field of MT and that they better align our evaluation with the current landscape. However, these changes are also likely responsible for some of the unexpected results observed this year, particularly for Task-1. We encourage further analysis of these results by the MT research community at large.

9 Ethical Considerations

The data for this shared task was generated, screened and human-annotated by the General Machine Translation Shared Task. We acknowledge inheriting any ethical limitations and concerns raised by their shared task. We do not foresee any additional ethical concerns.

10 Acknowledgments

Results for this shared task would not be possible without the tight collaboration with the organizers of the WMT25 General MT Shared Task. We thank them for their hard work and collaboration.

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

Chrysoula Zerva was funded by the UTTER project, supported by the European Union's Horizon Europe research and innovation programme via grant agreement 101070631, by the Portuguese Recovery and Resilience Plan through projects C645008882-00000055 (Center for Responsible AI) and UID/50008: Instituto de Telecomunicações and supported by an unrestricted gift from Google (Google Research Scholar).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88. Association for Computational Linguistics.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2024. Together we can: Multilingual automatic post-editing for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10800–10812. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2025. Giving the old a fresh spin: Quality estimation-assisted constrained decoding for automatic post-editing. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 914–925. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe, and Pushpak Bhattacharyya.

2023. Quality estimation-assisted automatic post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1686–1698. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929. Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10. Association for Computational Linguistics.

Giorgos Filandrianos, Orfeas Menis Mastromichalakis, Wafaa Mohammed, Giuseppe Attanasio, and Chrysoula Zerva. 2025. GAMBIT+: A Challenge Set for Evaluating Gender Bias in Machine Translation Quality Estimation Metrics. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Sami Ul Haq and Chinonso Cynthia Osuji. 2025. Long-context Reference-based MT Quality Estimation. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. CUNI and Phrase at WMT25 MT Evaluation Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2025. GEMBA V2: Ten Judgments Are Better Than One. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate Submissions to the WMT25 Evaluation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2024. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, pages 475–491. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, and Chi-kiu Lo. 2025. MSLC25: Metric Performance on Low-Quality Machine Translation, Empty Strings, and Language Variants. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025a. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025b. Preliminary ranking of WMT25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

Senyu Li, Felermino Dario Mario Ali, Jiayi Wang, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, Colin Cherry, and David Ifeoluwa Adelani. 2025a. Evaluating WMT 2025 Metrics Shared Task Submissions on the SSA-MTE African Challenge Set. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Senyu Li, Jiayi Wang, Felermino DMA Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025b. SSA-COMET: Do LLMs outperform learned metrics in evaluating MT for under-resourced african languages? *arXiv preprint arXiv:2506.04557*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513. Association for Computational Linguistics.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023a. Beyond correlation: Making sense of the score differences of new MT evaluation metrics. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 186–199. Asia-Pacific Association for Machine Translation.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023b. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799. Association for Computational Linguistics.

Yuanchang Luo, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, xiaoyu chen, and Hao Yang. 2025. HW-TSC's submissions to the WMT 2025 Segment-level quality score prediction Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. Ranked-COMET: Elevating a 2022 Baseline to a Top-5 Finish in the WMT 2025 QE Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Monishwaran Maheswaran, Marco Carini, Christian Federmann, and Tony Diaz. 2025. TASER: Translation assessment via systematic evaluation and reasoning. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Benjamin Marie. 2022. Yes, we need statistical significance testing. towardsai.net https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Eric W Noreen. 1989. *Computer-intensive methods for testing hypotheses*. Wiley New York.

Govardhan Padmanabhan. 2025. Can QE-informed (Re)Translation lead to Error Correction? In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645. Association for Computational Linguistics.

T. Robertson, F.T. Wright, and R. Dykstra. 1988. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892. Association for Computational Linguistics.

Prashant K. Sharma. 2025. Leveraging QE-based Explanations for Quality-Informed Corrections. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231. Association for Machine Translation in the Americas.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91. Association for Computational Linguistics.

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. CoDrift in WMT25 Metric Challenge Set Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234. Association for Computational Linguistics.

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 129–139. Association for Computational Linguistics.

Di Wu and Christof Monz. 2025. UvA-MT at WMT25 Evaluation Task: LLM Uncertainty as a Proxy for Translation Quality. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Brian Yan, Shuoyang Ding, Kuang-Da Wang, Siqi Ouyang, Oleksii Hrinchuk, Vitaly Lavrukhin, and Boris Ginsburg. 2025. Nvidia-Nemo's WMT 2025 Metrics Shared Task Submission. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and JinYeong Bak. 2025. Tagged Span Annotation for

Reasoning LLM-Based Translation Error Span Detection. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. *Preprint*, arXiv:1904.09675.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. How to select datapoints for efficient human evaluation of NLG models? *Preprint*, arXiv:2501.18251.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. COMET-poly: Machine Translation Metric Grounded in Other Candidates. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

A LLM Prompt for Task 1

Segment-level quality scores from the "LLM as a judge" submissions to Task 1 (Section 4.1.3) were prompted using the template below. Placeholders in curly braces indicate the name of the source language, name of the target language, source segment text, and target segment text.

```
Score the following translation from {source_lang} to
{target_lang} on a scale from 0 to 100, where a score of 0 means a
broken or poor translation; 33 indicates a flawed translation
with significant issues; 66 indicates a good translation with
only minor issues in grammar, fluency, or consistency; and 100
represents a perfect translation in both meaning and grammar.
Answer with only a whole number representing the score, and
nothing else.

{source_lang} source text:
{source_seg}
{target_lang} translation:
{target_seg}
```

B Complete Task 1 Results per Language Pair

Table 19 (part 1) and Table 20 (part 2) show the full detailed results of the segment-level quality score prediction task broken down by individual language pair. Correlations are computed using SPA at the system level and acc^*_{eq} at the segment level, against the "human1" gold-standard annotations, matching the approach taken in the summary results of Section 4.3.

Table 21 and Table 22 show a similar detailed per-language-pair breakdown of the results as above, except now using "human2" as the gold standard. Only language pairs annotated with ESA have this second human score; Japanese→Chinese and English→Korean are thus excluded from these tables.

Metric	cs-de		cs-uk		en-ar		en-bho		en-cs		en-et		en-is		en-it	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines																
YiSi-1	0.899 (1)	0.564 (1)	0.719 (5)	0.500 (6)	0.838 (2)	0.595 (4)	0.845 (2)	0.672 (1)	0.748 (3)	0.554 (1)	0.862 (1)	0.593 (3)	0.900 (3)	0.721 (3)	–	–
chrF	0.896 (2)	0.556 (2)	0.707 (5)	0.490 (7)	0.847 (2)	0.625 (3)	0.847 (2)	0.665 (2)	0.772 (3)	0.550 (3)	0.846 (1)	0.591 (4)	0.893 (3)	0.730 (2)	–	–
spBLEU	0.881 (3)	0.549 (3)	0.704 (5)	0.485 (8)	0.849 (2)	0.639 (2)	0.844 (2)	0.647 (3)	0.741 (4)	0.551 (3)	0.833 (2)	0.584 (6)	0.915 (2)	0.719 (3)	–	–
BERTScore	0.872 (3)	0.550 (3)	0.672 (6)	0.481 (9)	0.866 (1)	0.672 (1)	0.868 (1)	0.665 (2)	0.717 (5)	0.543 (4)	0.798 (3)	0.574 (7)	0.866 (4)	0.699 (5)	–	–
BLEU	0.873 (3)	0.540 (4)	0.661 (6)	0.475 (10)	0.853 (1)	0.628 (3)	0.872 (1)	0.644 (4)	0.734 (4)	0.539 (5)	0.794 (3)	0.567 (8)	0.879 (4)	0.693 (6)	–	–
COMET22	0.773 (6)	0.536 (5)	0.764 (4)	0.517 (4)	0.674 (5)	0.472 (7)	0.712 (5)	0.611 (6)	0.676 (6)	0.552 (3)	0.732 (4)	0.590 (5)	0.883 (4)	0.699 (5)	–	–
sentinel-cand	0.658 (8)	0.493 (10)	0.620 (6)	0.496 (6)	0.241 (13)	0.369 (13)	0.448 (12)	0.463 (15)	0.603 (8)	0.523 (7)	0.674 (6)	0.571 (7)	0.819 (6)	0.660 (9)	0.640 (7)	0.495 (5)
COMETKiwi22	0.622 (8)	0.493 (10)	0.572 (7)	0.456 (11)	0.140 (17)	0.369 (13)	0.473 (11)	0.468 (14)	0.538 (10)	0.497 (11)	0.595 (7)	0.542 (10)	0.761 (8)	0.651 (10)	0.612 (8)	0.482 (6)
sentinel-src	0.568 (9)	0.140 (23)	0.475 (8)	0.169 (26)	0.466 (8)	0.370 (12)	0.482 (11)	0.141 (28)	0.598 (8)	0.121 (27)	0.536 (9)	0.136 (25)	0.478 (10)	0.131 (30)	0.456 (10)	0.173 (22)
Primary																
GEMBA-v2	0.848 (4)	0.552 (3)	0.850 (1)	0.500 (6)	0.629 (6)	0.370 (12)	0.720 (5)	0.593 (7)	0.868 (1)	0.549 (3)	0.798 (3)	0.596 (3)	0.915 (2)	0.707 (4)	0.847 (1)	0.535 (1)
TASER-No-Ref	0.881 (2)	0.487 (11)	0.856 (1)	0.448 (12)	0.613 (7)	0.370 (12)	0.813 (3)	0.605 (6)	0.853 (1)	0.514 (8)	0.854 (1)	0.544 (10)	0.944 (1)	0.705 (4)	0.816 (2)	0.479 (6)
rankedCOMET	0.804 (5)	0.536 (4)	0.750 (4)	0.518 (3)	0.698 (4)	0.483 (6)	0.714 (5)	0.611 (5)	0.685 (5)	0.552 (2)	0.735 (4)	0.592 (4)	0.882 (4)	0.699 (5)	0.643 (7)	0.497 (5)
MetricX-25	0.773 (6)	0.539 (4)	0.823 (2)	0.528 (2)	0.364 (9)	0.370 (12)	0.647 (8)	0.552 (10)	0.744 (4)	0.551 (3)	0.751 (4)	0.597 (3)	0.858 (5)	0.692 (6)	0.728 (3)	0.539 (1)
mr7_2_1	0.853 (3)	0.471 (12)	0.847 (1)	0.450 (12)	0.247 (13)	0.370 (12)	0.659 (7)	0.523 (11)	0.828 (2)	0.450 (14)	0.764 (4)	0.485 (12)	0.832 (6)	0.607 (15)	0.826 (2)	0.431 (10)
SEGALE-QE	0.741 (6)	0.518 (6)	0.676 (5)	0.483 (8)	0.326 (10)	0.370 (12)	0.659 (7)	0.549 (10)	0.631 (7)	0.527 (6)	0.721 (5)	0.576 (7)	0.870 (4)	0.699 (5)	0.727 (3)	0.527 (2)
Polycand-2	0.694 (7)	0.503 (8)	0.687 (5)	0.498 (6)	0.250 (13)	0.369 (13)	0.521 (10)	0.480 (13)	0.621 (8)	0.528 (6)	0.733 (4)	0.572 (7)	0.878 (4)	0.680 (7)	0.682 (6)	0.502 (4)
Q_Relative-MQM	0.839 (4)	0.377 (16)	0.774 (4)	0.347 (17)	0.260 (12)	0.369 (13)	0.590 (9)	0.375 (18)	0.802 (3)	0.394 (17)	0.714 (5)	0.350 (17)	0.756 (8)	0.445 (21)	0.831 (1)	0.337 (16)
EnsembleSlick	0.718 (7)	0.484 (11)	0.646 (6)	0.457 (11)	0.332 (10)	0.369 (13)	0.538 (10)	0.468 (14)	0.629 (7)	0.496 (11)	0.728 (5)	0.541 (10)	0.833 (5)	0.625 (13)	0.715 (5)	0.493 (5)
hw-tsc	0.626 (8)	0.499 (9)	0.521 (7)	0.447 (12)	0.157 (16)	0.369 (13)	0.526 (10)	0.491 (12)	0.565 (9)	0.501 (10)	0.575 (8)	0.542 (10)	0.843 (5)	0.657 (9)	0.666 (6)	0.494 (5)
UvA-MT	0.557 (9)	0.512 (7)	0.518 (7)	0.462 (11)	0.349 (9)	0.369 (13)	0.386 (13)	0.441 (16)	0.525 (10)	0.497 (10)	0.317 (10)	0.447 (14)	0.502 (10)	0.503 (18)	0.552 (9)	0.464 (8)
Roberta-LS	–	–	–	–	–	–	–	–	0.643 (7)	0.509 (9)	–	–	–	–	–	–
Secondary																
TASER-Ref	0.917 (1)	0.553 (2)	0.874 (1)	0.512 (5)	0.710 (4)	0.432 (9)	0.797 (3)	0.651 (3)	0.872 (1)	0.557 (1)	0.867 (1)	0.610 (1)	0.940 (1)	0.736 (1)	0.851 (1)	0.524 (2)
MetricX-25-Ref	0.811 (5)	0.557 (2)	0.822 (3)	0.535 (1)	0.322 (11)	0.370 (12)	0.650 (7)	0.560 (9)	0.762 (3)	0.558 (1)	0.772 (3)	0.601 (2)	0.866 (4)	0.702 (5)	–	–
baseCOMET	0.773 (6)	0.536 (4)	0.764 (4)	0.518 (3)	0.674 (6)	0.472 (7)	0.712 (5)	0.611 (5)	0.676 (6)	0.553 (2)	0.732 (4)	0.592 (4)	0.883 (4)	0.699 (5)	0.643 (7)	0.497 (5)
MetricX-25-QE	0.727 (7)	0.520 (6)	0.752 (4)	0.511 (5)	0.338 (10)	0.370 (12)	0.661 (7)	0.565 (8)	0.692 (5)	0.541 (4)	0.732 (4)	0.589 (5)	0.833 (6)	0.678 (7)	0.710 (5)	0.539 (1)
mr6	0.809 (5)	0.470 (12)	0.846 (2)	0.438 (13)	0.285 (12)	0.370 (12)	0.655 (7)	0.518 (11)	0.795 (3)	0.437 (15)	0.759 (4)	0.463 (13)	0.824 (6)	0.599 (16)	0.798 (2)	0.424 (11)
Q_MQM	0.843 (4)	0.385 (15)	0.768 (4)	0.358 (16)	0.264 (12)	0.369 (13)	0.586 (9)	0.393 (17)	0.804 (2)	0.406 (16)	0.709 (5)	0.363 (16)	0.753 (8)	0.460 (20)	0.829 (1)	0.341 (15)
Polyic-3	0.665 (8)	0.497 (9)	0.700 (5)	0.500 (6)	0.232 (14)	0.370 (12)	0.464 (11)	0.465 (14)	0.608 (8)	0.521 (7)	0.690 (6)	0.572 (7)	0.839 (5)	0.672 (8)	0.658 (7)	0.501 (4)
AutoLQA	0.698 (7)	0.387 (15)	0.652 (6)	0.367 (15)	0.625 (6)	0.369 (13)	0.644 (8)	0.394 (17)	0.714 (5)	0.384 (18)	0.801 (3)	0.386 (15)	0.869 (4)	0.434 (22)	0.758 (3)	0.418 (11)
Polycand-1	0.679 (7)	0.499 (9)	0.645 (6)	0.490 (7)	0.241 (13)	0.378 (11)	0.484 (11)	0.465 (14)	0.613 (8)	0.524 (7)	0.700 (5)	0.565 (8)	0.839 (5)	0.670 (8)	0.666 (6)	0.497 (5)
CollabPlus	0.830 (4)	0.517 (6)	0.657 (6)	0.474 (10)	0.325 (10)	0.370 (12)	0.526 (10)	0.467 (14)	0.617 (8)	0.509 (9)	0.717 (5)	0.550 (9)	0.793 (7)	0.618 (14)	0.693 (5)	0.507 (3)
CollabSlick	0.752 (6)	0.506 (8)	0.694 (5)	0.474 (10)	0.308 (11)	0.369 (13)	0.544 (10)	0.480 (13)	0.649 (6)	0.517 (8)	0.733 (4)	0.551 (9)	0.827 (6)	0.642 (11)	0.723 (4)	0.507 (3)
hw-tsc-max	0.588 (9)	0.484 (11)	0.436 (8)	0.437 (13)	0.178 (15)	0.369 (13)	0.484 (11)	0.464 (15)	0.581 (9)	0.509 (9)	0.618 (7)	0.550 (9)	0.767 (8)	0.631 (12)	0.616 (8)	0.478 (7)
hw-tsc-base	0.588 (9)	0.484 (11)	0.436 (8)	0.437 (13)	0.145 (17)	0.369 (13)	0.484 (11)	0.464 (15)	0.541 (10)	0.491 (12)	0.531 (9)	0.521 (11)	0.767 (8)	0.631 (12)	0.616 (8)	0.478 (7)
long-context	–	–	–	–	–	–	–	–	0.639 (7)	0.510 (9)	–	–	–	–	–	–
roberta-multi	–	–	–	–	–	–	–	–	0.624 (8)	0.502 (10)	–	–	–	–	–	–
LLM-as-a-judge																
GPT-4_1	0.899 (2)	0.464 (13)	0.847 (2)	0.404 (14)	0.868 (1)	0.531 (5)	0.777 (4)	0.563 (8)	0.869 (1)	0.469 (13)	0.840 (2)	0.522 (11)	0.929 (2)	0.660 (9)	0.840 (1)	0.450 (9)
CommandA	0.859 (3)	0.406 (14)	0.833 (2)	0.354 (16)	0.737 (3)	0.439 (8)	0.694 (6)	0.341 (20)	0.871 (1)	0.377 (19)	0.783 (3)	0.350 (17)	0.808 (7)	0.429 (22)	0.831 (1)	0.397 (12)
Claude-4	0.876 (3)	0.348 (18)	0.839 (2)	0.291 (21)	0.729 (3)	0.422 (10)	0.774 (4)	0.348 (19)	0.880 (1)	0.281 (22)	0.791 (3)	0.298 (19)	0.907 (3)	0.519 (17)	0.836 (1)	0.283 (17)
DeepSeek-V3	0.857 (3)	0.364 (17)	0.854 (1)	0.331 (18)	0.554 (8)	0.369 (13)	0.639 (8)	0.379 (18)	0.848 (2)	0.307 (21)	0.846 (1)	0.330 (18)	0.862 (4)	0.473 (19)	0.826 (2)	0.338 (15)
Qwen3-235B	0.837 (4)	0.388 (15)	0.845 (2)	0.327 (19)	0.530 (8)	0.369 (13)	0.611 (9)	0.287 (22)	0.879 (1)	0.353 (20)	0.807 (3)	0.329 (18)	0.837 (5)	0.425 (23)	0.834 (1)	0.380 (13)
Qwen2_5-7B	0.721 (7)	0.362 (17)	0.657 (6)	0.320 (20)	0.339 (9)	0.369 (13)	0.535 (10)	0.313 (21)	0.779 (3)	0.307 (21)	0.541 (8)	0.292 (20)	0.659 (9)	0.378 (24)	0.817 (2)	0.368 (14)
AyaExpanse-32B	0.791 (5)	0.278 (19)	0.826 (2)	0.293 (21)	0.542 (8)	0.369 (13)	0.635 (8)	0.269 (23)	0.834 (2)	0.218 (25)	0.631 (7)	0.153 (24)	0.649 (9)	0.215 (28)	0.822 (2)	0.278 (18)
Llama-3_1-8B	0.747 (6)	0.267 (20)	0.705 (5)	0.250 (23)	0.202 (15)	0.369 (13)	0.577 (9)	0.197 (25)	0.738 (4)	0.223 (24)	0.629 (7)	0.252 (21)	0.740 (8)	0.224 (27)	0.778 (2)	0.239 (20)
Llama-4-Maverick	0.817 (4)	0.142 (23)	0.825 (2)	0.165 (27)	0.587 (7)	0.369 (13)	0.657 (7)	0.186 (26)	0.655 (6)	0.053 (28)	0.751 (4)	0.056 (27)	0.783 (7)	0.244 (25)	0.736 (3)	0.113 (23)
CommandR7B	0.769 (6)	0.218 (21)	0.708 (5)	0.214 (24)	0.171 (15)	0.369 (13)	0.549 (10)	0.181 (27)	0.682 (5)	0.186 (26)	0.465 (9)	0.195 (23)	0.442 (11)	0.186 (29)	0.803 (2)	0.258 (19)
Mistral-7B	0.639 (8)	0.282 (19)	0.560 (7)	0.276 (22)	0.207 (14)	0.369 (13)	0.476 (11)	0.221 (24)	0.618 (8)	0.233 (23)	0.487 (9)	0.222 (22)	0.463 (10)	0.238 (26)	0.630 (7)	0.258 (19)
AyaExpanse-8B	0.760 (6)	0.188 (22)	0.658 (6)	0.196 (25)	0.336 (10)	0.369 (13)	0.487 (11)	0.200 (25)	0.703 (5)	0.124 (27)	0.455 (9)	0.111 (26)	0.440 (11)	0.114 (31)	0.716 (4)	0.185 (21)

Table 19: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human1" as the gold standard. Metrics are ordered by category. Part 1 of 2.

Metric	en-ja		en-ko		en-mas		en-ru		en-sr		en-uk		en-zh		ja-zh	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines																
YiSi-1	0.836 (1)	0.534 (1)	0.839 (3)	0.470 (4)	–	–	0.557 (6)	0.503 (8)	0.864 (2)	0.604 (1)	0.645 (5)	0.515 (6)	0.643 (7)	0.486 (5)	0.877 (2)	0.503 (1)
chrF	0.813 (1)	0.525 (2)	0.852 (2)	0.465 (5)	–	–	0.542 (6)	0.492 (10)	0.840 (4)	0.589 (3)	0.677 (4)	0.520 (5)	0.658 (6)	0.479 (6)	0.857 (2)	0.487 (3)
spBLEU	0.808 (1)	0.520 (3)	0.818 (3)	0.464 (5)	–	–	0.597 (5)	0.504 (8)	0.858 (3)	0.590 (3)	0.648 (5)	0.510 (7)	0.638 (7)	0.472 (7)	0.837 (3)	0.486 (3)
BERTScore	0.818 (1)	0.521 (2)	0.805 (4)	0.466 (5)	–	–	0.583 (5)	0.498 (9)	0.838 (4)	0.596 (2)	0.604 (7)	0.505 (8)	0.621 (8)	0.482 (6)	0.853 (3)	0.491 (2)
BLEU	0.803 (1)	0.512 (4)	0.827 (3)	0.466 (5)	–	–	0.540 (6)	0.489 (10)	0.844 (4)	0.582 (4)	0.638 (6)	0.515 (6)	0.632 (7)	0.475 (7)	0.837 (3)	0.482 (4)
COMET22	0.462 (9)	0.493 (6)	0.806 (4)	0.474 (3)	–	–	0.557 (6)	0.514 (7)	0.784 (6)	0.582 (4)	0.689 (4)	0.540 (2)	0.652 (6)	0.496 (4)	0.766 (4)	0.465 (5)
sentinel-cand	0.322 (11)	0.444 (10)	0.638 (6)	0.457 (7)	0.545 (6)	0.489 (6)	0.551 (6)	0.506 (8)	0.731 (9)	0.546 (8)	0.638 (6)	0.526 (5)	0.614 (8)	0.470 (7)	0.447 (8)	0.388 (13)
COMETKiwi22	0.470 (9)	0.466 (8)	0.507 (8)	0.459 (7)	0.422 (10)	0.489 (7)	0.448 (7)	0.470 (12)	0.736 (8)	0.554 (7)	0.583 (7)	0.503 (8)	0.477 (9)	0.454 (9)	0.448 (8)	0.389 (13)
sentinel-src	0.530 (7)	0.157 (22)	0.523 (7)	0.460 (7)	0.546 (6)	0.493 (5)	0.598 (5)	0.143 (24)	0.362 (13)	0.165 (22)	0.536 (7)	0.149 (24)	0.513 (9)	0.179 (23)	0.458 (8)	0.235 (22)
Primary																
GEMBA-v2	0.712 (3)	0.512 (4)	0.901 (1)	0.471 (4)	0.682 (1)	0.489 (6)	0.775 (2)	0.551 (1)	0.875 (2)	0.587 (3)	0.773 (2)	0.532 (4)	0.832 (2)	0.487 (5)	0.858 (2)	0.460 (6)
TASER-No-Ref	0.736 (2)	0.447 (10)	0.915 (1)	0.466 (5)	0.604 (3)	0.507 (4)	0.812 (1)	0.496 (9)	0.881 (1)	0.559 (6)	0.786 (1)	0.478 (11)	0.807 (2)	0.443 (10)	0.912 (1)	0.423 (8)
rankedCOMET	0.458 (9)	0.493 (5)	0.829 (3)	0.458 (7)	0.540 (6)	0.489 (7)	0.568 (5)	0.515 (6)	0.768 (7)	0.582 (4)	0.697 (4)	0.539 (3)	0.665 (6)	0.497 (3)	0.775 (4)	0.465 (6)
MetricX-25	0.522 (8)	0.496 (5)	0.852 (2)	0.456 (9)	0.437 (10)	0.489 (6)	0.680 (3)	0.545 (2)	0.812 (5)	0.583 (4)	0.693 (4)	0.543 (2)	0.716 (5)	0.504 (2)	0.719 (5)	0.470 (5)
mr7_2_1	0.651 (5)	0.443 (10)	0.878 (1)	0.469 (4)	0.579 (4)	0.489 (7)	0.774 (2)	0.459 (13)	0.805 (5)	0.518 (11)	0.802 (1)	0.438 (12)	0.857 (1)	0.436 (11)	0.846 (3)	0.427 (8)
SEGALE-QE	0.543 (7)	0.479 (7)	0.736 (5)	0.459 (7)	0.537 (6)	0.489 (7)	0.565 (6)	0.510 (7)	0.722 (9)	0.529 (10)	0.717 (3)	0.525 (5)	0.649 (7)	0.488 (5)	0.603 (6)	0.412 (10)
Polycand-2	0.364 (11)	0.467 (8)	0.732 (5)	0.466 (5)	0.516 (7)	0.489 (7)	0.607 (4)	0.526 (4)	0.771 (7)	0.562 (6)	0.648 (5)	0.531 (4)	0.644 (7)	0.482 (6)	0.612 (6)	0.425 (8)
Q_Relative-MQM	0.732 (2)	0.386 (14)	0.811 (3)	0.460 (7)	0.542 (6)	0.489 (7)	0.734 (2)	0.368 (17)	0.860 (2)	0.450 (14)	0.764 (2)	0.354 (16)	0.831 (2)	0.381 (14)	0.848 (3)	0.414 (10)
EnsembleSlick	0.447 (10)	0.472 (8)	0.669 (6)	0.467 (5)	0.504 (7)	0.489 (6)	0.566 (6)	0.513 (7)	0.401 (12)	0.430 (15)	0.640 (6)	0.494 (9)	0.650 (7)	0.471 (7)	0.595 (7)	0.399 (12)
hw-tsc	0.527 (8)	0.476 (7)	0.622 (6)	0.456 (9)	0.502 (7)	0.489 (7)	0.446 (7)	0.487 (10)	0.759 (7)	0.554 (7)	0.633 (6)	0.513 (6)	0.476 (9)	0.468 (8)	0.526 (7)	0.403 (11)
UvA-MT	0.601 (6)	0.490 (6)	0.458 (9)	0.457 (8)	0.435 (10)	0.489 (6)	0.445 (7)	0.468 (12)	0.658 (10)	0.524 (10)	0.409 (9)	0.487 (10)	0.376 (11)	0.440 (10)	0.417 (8)	0.391 (13)
Roberta-LS	0.491 (9)	0.435 (12)	–	–	–	–	–	–	–	–	–	–	0.661 (6)	0.432 (11)	–	–
Secondary																
TASER-Ref	0.725 (3)	0.494 (5)	0.901 (1)	0.499 (1)	0.624 (2)	0.532 (3)	0.800 (1)	0.536 (3)	0.890 (1)	0.594 (2)	0.810 (1)	0.525 (5)	0.825 (2)	0.484 (5)	0.920 (1)	0.500 (1)
MetricX-25-Ref	0.559 (6)	0.509 (4)	0.851 (2)	0.458 (7)	–	–	0.686 (3)	0.556 (1)	0.824 (4)	0.588 (3)	0.735 (3)	0.550 (1)	0.756 (4)	0.513 (1)	0.768 (4)	0.490 (2)
baseCOMET	0.462 (9)	0.493 (5)	0.806 (4)	0.474 (3)	0.567 (5)	0.489 (7)	0.557 (6)	0.515 (6)	0.784 (6)	0.582 (4)	0.689 (4)	0.540 (2)	0.652 (7)	0.497 (3)	0.766 (4)	0.465 (6)
MetricX-25-QE	0.496 (9)	0.489 (6)	0.822 (3)	0.457 (8)	0.479 (8)	0.489 (7)	0.633 (4)	0.540 (3)	0.793 (6)	0.568 (5)	0.674 (4)	0.542 (2)	0.709 (5)	0.505 (2)	0.675 (6)	0.458 (7)
mr6	0.630 (5)	0.435 (11)	0.861 (2)	0.477 (2)	0.566 (5)	0.489 (7)	0.716 (3)	0.433 (15)	0.795 (5)	0.517 (11)	0.758 (2)	0.410 (13)	0.797 (3)	0.417 (12)	0.810 (3)	0.422 (8)
Q_MQM	0.732 (2)	0.399 (13)	0.819 (3)	0.463 (6)	0.651 (2)	0.489 (7)	0.733 (2)	0.372 (16)	0.858 (3)	0.459 (13)	0.758 (2)	0.360 (15)	0.833 (1)	0.387 (13)	0.841 (3)	0.417 (9)
Polyic-3	0.347 (11)	0.460 (9)	0.711 (5)	0.478 (2)	0.482 (8)	0.489 (7)	0.590 (5)	0.520 (5)	0.762 (7)	0.549 (8)	0.639 (6)	0.529 (4)	0.638 (7)	0.482 (6)	0.612 (6)	0.420 (9)
AutoLQA	0.536 (7)	0.391 (14)	0.762 (4)	0.481 (2)	0.436 (10)	0.489 (7)	0.702 (3)	0.441 (14)	0.808 (5)	0.446 (14)	0.724 (3)	0.393 (14)	0.735 (4)	0.347 (16)	0.629 (6)	0.357 (15)
Polycand-1	0.340 (11)	0.469 (8)	0.713 (5)	0.461 (6)	0.491 (8)	0.489 (7)	0.598 (5)	0.520 (5)	0.751 (8)	0.553 (7)	0.637 (6)	0.525 (5)	0.642 (7)	0.479 (6)	0.605 (6)	0.419 (9)
CollabPlus	0.419 (10)	0.470 (8)	0.697 (5)	0.464 (5)	0.422 (10)	0.489 (6)	0.555 (6)	0.510 (7)	0.415 (12)	0.428 (15)	0.668 (4)	0.512 (6)	0.651 (7)	0.478 (6)	0.697 (5)	0.419 (9)
CollabSlick	0.454 (9)	0.480 (7)	0.687 (5)	0.467 (4)	0.496 (7)	0.489 (6)	0.573 (5)	0.521 (4)	0.387 (13)	0.423 (16)	0.647 (5)	0.502 (8)	0.662 (6)	0.479 (6)	0.613 (6)	0.405 (11)
hw-tsc-max	0.549 (7)	0.474 (7)	0.572 (7)	0.456 (9)	0.466 (9)	0.489 (7)	0.441 (7)	0.475 (11)	0.711 (9)	0.542 (9)	0.608 (7)	0.503 (8)	0.443 (10)	0.467 (8)	0.521 (7)	0.396 (12)
hw-tsc-base	0.549 (7)	0.474 (7)	0.491 (8)	0.456 (9)	0.466 (9)	0.489 (7)	0.441 (7)	0.475 (11)	0.711 (9)	0.542 (9)	0.608 (7)	0.503 (8)	0.443 (10)	0.467 (8)	0.521 (7)	0.396 (12)
long-context	0.445 (10)	0.447 (10)	–	–	–	–	–	–	–	–	–	–	0.686 (5)	0.431 (11)	–	–
roberta-multi	0.477 (9)	0.440 (11)	–	–	–	–	–	–	–	–	–	–	0.646 (7)	0.437 (10)	–	–
LLM-as-a-judge																
GPT-4_1	0.761 (2)	0.454 (9)	0.877 (2)	0.503 (1)	0.610 (3)	0.543 (2)	0.816 (1)	0.458 (13)	0.888 (1)	0.506 (12)	0.781 (2)	0.434 (12)	0.798 (3)	0.415 (12)	0.940 (1)	0.464 (6)
CommandA	0.704 (4)	0.327 (16)	0.890 (1)	0.471 (4)	0.590 (4)	0.489 (7)	0.815 (1)	0.364 (17)	0.832 (4)	0.429 (15)	0.800 (1)	0.344 (17)	0.833 (2)	0.336 (17)	0.913 (1)	0.375 (14)
Claude-4	0.746 (2)	0.304 (17)	0.915 (1)	0.476 (3)	0.622 (2)	0.555 (1)	0.769 (2)	0.283 (20)	0.877 (2)	0.418 (16)	0.775 (2)	0.253 (20)	0.847 (1)	0.302 (19)	0.934 (1)	0.391 (13)
DeepSeek-V3	0.666 (5)	0.282 (18)	0.904 (1)	0.476 (3)	0.606 (3)	0.489 (7)	0.755 (2)	0.330 (19)	0.853 (3)	0.451 (14)	0.806 (1)	0.311 (19)	0.845 (1)	0.355 (15)	0.875 (2)	0.389 (13)
Qwen3-235B	0.699 (4)	0.337 (15)	0.888 (1)	0.472 (3)	0.589 (4)	0.489 (7)	0.799 (1)	0.361 (18)	0.841 (4)	0.415 (17)	0.809 (1)	0.327 (18)	0.863 (1)	0.358 (15)	0.812 (3)	0.398 (12)
Qwen2_5-7B	0.589 (6)	0.280 (18)	0.802 (4)	0.470 (4)	0.575 (5)	0.489 (7)	0.770 (2)	0.363 (17)	0.742 (8)	0.371 (18)	0.638 (6)	0.343 (17)	0.785 (3)	0.327 (18)	0.777 (4)	0.361 (15)
AyaExpanse-32B	0.626 (5)	0.225 (21)	0.896 (1)	0.456 (8)	0.582 (4)	0.489 (7)	0.744 (2)	0.227 (23)	0.690 (10)	0.274 (19)	0.766 (2)	0.219 (22)	0.821 (2)	0.237 (22)	0.794 (4)	0.331 (16)
Llama-3_1-8B	0.594 (6)	0.226 (21)	0.801 (4)	0.456 (9)	0.546 (6)	0.489 (7)	0.714 (3)	0.248 (21)	0.772 (6)	0.261 (20)	0.619 (6)	0.205 (23)	0.813 (2)	0.277 (20)	0.630 (6)	0.310 (18)
Llama-4-Maverick	0.652 (5)	0.098 (24)	0.818 (3)	0.463 (5)	0.602 (3)	0.489 (7)	0.645 (3)	0.070 (26)	0.736 (8)	0.180 (21)	0.650 (4)	0.064 (26)	0.800 (2)	0.117 (24)	0.841 (3)	0.261 (21)
CommandR7B	0.615 (5)	0.253 (19)	0.761 (4)	0.472 (3)	0.566 (5)	0.489 (7)	0.571 (5)	0.239 (22)	0.371 (13)	0.167 (22)	0.500 (8)	0.217 (22)	0.682 (5)	0.264 (21)	0.674 (6)	0.317 (17)
Mistral-7B	0.443 (10)	0.247 (20)	0.596 (6)	0.456 (9)	0.559 (5)	0.489 (7)	0.533 (6)	0.251 (21)	0.613 (11)	0.271 (19)	0.554 (7)	0.231 (21)	0.623 (7)	0.234 (22)	0.571 (7)	0.302 (19)
AyaExpanse-8B	0.558 (6)	0.131 (23)	0.729 (5)	0.456 (9)	0.502 (7)	0.489 (7)	0.633 (3)	0.126 (25)	0.422 (12)	0.182 (21)	0.509 (8)	0.123 (25)	0.670 (5)	0.178 (23)	0.703 (5)	0.288 (20)

Table 20: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human1" as the gold standard. Metrics are ordered by category. Part 2 of 2.

Metric	cs-de		cs-uk		en-ar		en-bho		en-cs		en-et		en-is	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines														
YiSi-1	0.863 (1)	0.506 (1)	0.696 (5)	0.465 (5)	0.830 (2)	0.599 (5)	0.903 (2)	0.667 (1)	0.790 (3)	0.567 (5)	0.853 (1)	0.623 (2)	0.886 (3)	0.706 (2)
chrF	0.856 (1)	0.501 (2)	0.684 (5)	0.455 (6)	0.839 (2)	0.628 (4)	0.858 (3)	0.641 (3)	0.815 (2)	0.567 (5)	0.853 (1)	0.621 (2)	0.867 (4)	0.705 (2)
spBLEU	0.842 (2)	0.497 (3)	0.676 (5)	0.457 (6)	0.852 (1)	0.647 (2)	0.902 (2)	0.635 (4)	0.789 (3)	0.556 (6)	0.864 (1)	0.622 (2)	0.895 (3)	0.698 (3)
BERTScore	0.840 (2)	0.496 (3)	0.644 (6)	0.448 (7)	0.870 (1)	0.678 (1)	0.922 (1)	0.650 (2)	0.760 (4)	0.556 (6)	0.827 (2)	0.606 (4)	0.856 (4)	0.681 (5)
BLEU	0.837 (2)	0.493 (4)	0.635 (6)	0.441 (8)	0.866 (1)	0.636 (3)	0.913 (1)	0.631 (5)	0.786 (3)	0.554 (6)	0.817 (2)	0.591 (6)	0.866 (4)	0.675 (6)
COMET22	0.754 (5)	0.495 (3)	0.774 (3)	0.478 (3)	0.628 (6)	0.469 (8)	0.772 (5)	0.604 (7)	0.684 (5)	0.576 (4)	0.691 (7)	0.607 (4)	0.869 (4)	0.690 (4)
sentinel-cand	0.650 (7)	0.460 (8)	0.622 (6)	0.463 (5)	0.195 (14)	0.362 (14)	0.508 (12)	0.465 (17)	0.598 (9)	0.539 (8)	0.643 (9)	0.575 (8)	0.803 (6)	0.652 (9)
COMETKiwi22	0.606 (8)	0.441 (10)	0.587 (6)	0.420 (11)	0.102 (17)	0.362 (14)	0.530 (11)	0.469 (17)	0.531 (10)	0.511 (11)	0.612 (9)	0.552 (11)	0.741 (7)	0.640 (11)
sentinel-src	0.544 (9)	0.196 (22)	0.501 (8)	0.208 (25)	0.464 (9)	0.363 (12)	0.515 (11)	0.144 (30)	0.563 (10)	0.122 (25)	0.517 (11)	0.131 (28)	0.474 (9)	0.129 (29)
Primary														
GEMBA-v2	0.818 (3)	0.494 (3)	0.878 (1)	0.471 (4)	0.624 (7)	0.365 (12)	0.763 (5)	0.589 (8)	0.839 (2)	0.579 (2)	0.846 (1)	0.617 (3)	0.916 (2)	0.690 (4)
TASER-No-Ref	0.845 (2)	0.458 (8)	0.867 (1)	0.424 (10)	0.605 (7)	0.362 (13)	0.856 (3)	0.588 (8)	0.841 (2)	0.524 (10)	0.872 (1)	0.569 (9)	0.945 (1)	0.680 (5)
rankedCOMET	0.783 (4)	0.495 (3)	0.747 (4)	0.478 (3)	0.652 (5)	0.482 (7)	0.772 (5)	0.605 (6)	0.694 (5)	0.577 (3)	0.699 (5)	0.606 (5)	0.866 (4)	0.690 (4)
MetricX-25	0.762 (5)	0.493 (3)	0.840 (2)	0.483 (2)	0.330 (10)	0.362 (13)	0.711 (7)	0.551 (11)	0.738 (4)	0.583 (2)	0.725 (5)	0.616 (3)	0.878 (4)	0.681 (5)
mr7_2_1	0.801 (3)	0.445 (10)	0.879 (1)	0.435 (9)	0.217 (14)	0.362 (13)	0.720 (6)	0.528 (13)	0.842 (1)	0.459 (14)	0.743 (4)	0.503 (14)	0.856 (4)	0.610 (15)
SEGALE-QE	0.755 (5)	0.474 (6)	0.695 (5)	0.454 (6)	0.286 (12)	0.362 (14)	0.716 (6)	0.538 (12)	0.640 (7)	0.545 (7)	0.686 (7)	0.581 (7)	0.884 (3)	0.679 (5)
Polycand-2	0.684 (7)	0.465 (7)	0.701 (4)	0.457 (6)	0.203 (14)	0.362 (14)	0.568 (10)	0.480 (16)	0.626 (8)	0.541 (7)	0.700 (5)	0.574 (8)	0.856 (4)	0.662 (8)
Q_Relative-MQM	0.803 (3)	0.386 (13)	0.795 (3)	0.368 (15)	0.254 (13)	0.362 (14)	0.644 (9)	0.393 (19)	0.812 (2)	0.414 (16)	0.728 (4)	0.365 (19)	0.777 (7)	0.437 (21)
EnsembleSlick	0.688 (7)	0.455 (9)	0.664 (5)	0.425 (10)	0.290 (11)	0.362 (14)	0.603 (9)	0.468 (17)	0.629 (8)	0.513 (11)	0.692 (6)	0.541 (12)	0.824 (5)	0.613 (14)
hw-tsc	0.610 (8)	0.452 (9)	0.540 (7)	0.411 (12)	0.115 (17)	0.362 (14)	0.586 (10)	0.498 (15)	0.570 (10)	0.522 (10)	0.583 (10)	0.554 (11)	0.840 (5)	0.646 (10)
UvA-MT	0.551 (9)	0.458 (8)	0.541 (7)	0.428 (10)	0.372 (10)	0.362 (14)	0.445 (13)	0.465 (17)	0.505 (11)	0.508 (11)	0.332 (13)	0.451 (16)	0.531 (9)	0.513 (17)
Roberta-LS	–	–	–	–	–	–	–	–	0.654 (7)	0.524 (10)	–	–	–	–
Secondary														
TASER-Ref	0.886 (1)	0.504 (2)	0.877 (1)	0.487 (2)	0.698 (4)	0.437 (10)	0.852 (3)	0.651 (2)	0.859 (1)	0.582 (2)	0.849 (1)	0.642 (1)	0.942 (1)	0.722 (1)
MetricX-25-Ref	0.793 (4)	0.510 (1)	0.844 (2)	0.494 (1)	0.289 (12)	0.362 (13)	0.699 (7)	0.550 (11)	0.754 (4)	0.593 (1)	0.742 (4)	0.621 (2)	0.873 (4)	0.686 (4)
baseCOMET	0.754 (5)	0.495 (3)	0.774 (3)	0.478 (3)	0.628 (7)	0.469 (8)	0.772 (5)	0.605 (6)	0.684 (6)	0.577 (3)	0.691 (6)	0.607 (4)	0.869 (4)	0.690 (4)
MetricX-25-QE	0.717 (6)	0.482 (5)	0.764 (3)	0.473 (3)	0.303 (11)	0.362 (13)	0.726 (6)	0.556 (10)	0.678 (6)	0.570 (5)	0.696 (6)	0.602 (5)	0.845 (5)	0.667 (7)
mr6	0.779 (4)	0.452 (9)	0.872 (1)	0.434 (9)	0.256 (13)	0.362 (14)	0.723 (6)	0.521 (14)	0.790 (3)	0.458 (14)	0.741 (4)	0.478 (15)	0.839 (5)	0.592 (16)
Q_MQM	0.805 (3)	0.394 (12)	0.787 (3)	0.379 (14)	0.258 (13)	0.362 (14)	0.642 (9)	0.407 (18)	0.812 (2)	0.422 (15)	0.723 (5)	0.377 (18)	0.777 (7)	0.451 (20)
Polyic-3	0.666 (7)	0.458 (8)	0.706 (4)	0.458 (6)	0.184 (15)	0.362 (14)	0.526 (11)	0.476 (16)	0.604 (9)	0.533 (9)	0.664 (8)	0.574 (8)	0.820 (6)	0.653 (9)
AutoLQA	0.711 (6)	0.381 (13)	0.664 (5)	0.378 (14)	0.589 (7)	0.362 (13)	0.684 (7)	0.397 (19)	0.718 (4)	0.394 (17)	0.755 (3)	0.386 (17)	0.884 (3)	0.433 (21)
Polycand-1	0.670 (7)	0.463 (7)	0.665 (5)	0.453 (7)	0.198 (14)	0.365 (12)	0.526 (11)	0.468 (17)	0.611 (8)	0.540 (7)	0.664 (8)	0.569 (9)	0.821 (6)	0.651 (9)
CollabPlus	0.790 (4)	0.480 (5)	0.669 (5)	0.438 (8)	0.282 (12)	0.362 (13)	0.592 (10)	0.466 (17)	0.624 (8)	0.537 (8)	0.682 (7)	0.558 (10)	0.777 (7)	0.615 (14)
CollabSlick	0.717 (6)	0.473 (6)	0.712 (4)	0.439 (8)	0.267 (12)	0.362 (14)	0.607 (9)	0.480 (16)	0.650 (7)	0.539 (8)	0.696 (5)	0.553 (11)	0.822 (5)	0.628 (12)
hw-tsc-max	0.580 (9)	0.438 (11)	0.450 (8)	0.405 (13)	0.138 (16)	0.362 (14)	0.546 (11)	0.479 (16)	0.592 (9)	0.532 (9)	0.614 (9)	0.563 (10)	0.753 (7)	0.621 (13)
hw-tsc-base	0.580 (9)	0.438 (11)	0.450 (8)	0.405 (13)	0.109 (17)	0.362 (14)	0.546 (11)	0.479 (16)	0.529 (11)	0.503 (12)	0.564 (11)	0.529 (13)	0.753 (7)	0.621 (13)
long-context	–	–	–	–	–	–	–	–	0.647 (7)	0.533 (9)	–	–	–	–
roberta-multi	–	–	–	–	–	–	–	–	0.633 (8)	0.526 (10)	–	–	–	–
LLM-as-a-judge														
GPT-4_1	0.845 (2)	0.435 (11)	0.852 (2)	0.405 (13)	0.847 (2)	0.547 (6)	0.823 (4)	0.570 (9)	0.868 (1)	0.491 (13)	0.851 (1)	0.549 (11)	0.944 (1)	0.645 (10)
CommandA	0.845 (2)	0.397 (12)	0.854 (2)	0.380 (14)	0.723 (3)	0.449 (9)	0.765 (5)	0.351 (22)	0.868 (1)	0.390 (17)	0.755 (4)	0.367 (19)	0.826 (5)	0.425 (22)
Claude-4	0.856 (2)	0.351 (17)	0.852 (2)	0.338 (17)	0.702 (3)	0.417 (11)	0.829 (4)	0.364 (21)	0.871 (1)	0.289 (20)	0.775 (3)	0.315 (22)	0.918 (2)	0.506 (18)
DeepSeek-V3	0.824 (3)	0.375 (15)	0.873 (1)	0.362 (16)	0.531 (8)	0.362 (14)	0.711 (7)	0.378 (20)	0.870 (1)	0.325 (19)	0.783 (3)	0.347 (20)	0.857 (4)	0.465 (19)
Qwen3-235B	0.799 (4)	0.379 (14)	0.862 (1)	0.343 (17)	0.502 (9)	0.362 (14)	0.683 (8)	0.302 (24)	0.870 (1)	0.367 (18)	0.767 (3)	0.336 (21)	0.861 (4)	0.421 (22)
Qwen2_5-7B	0.704 (6)	0.360 (16)	0.667 (5)	0.336 (18)	0.313 (11)	0.362 (13)	0.599 (10)	0.322 (23)	0.782 (3)	0.322 (19)	0.530 (11)	0.290 (23)	0.670 (8)	0.365 (23)
AyaExpanse-32B	0.777 (4)	0.300 (18)	0.863 (1)	0.323 (19)	0.517 (8)	0.362 (13)	0.689 (7)	0.276 (25)	0.850 (1)	0.225 (23)	0.588 (10)	0.157 (27)	0.687 (8)	0.208 (27)
Llama-3_1-8B	0.718 (5)	0.293 (19)	0.727 (4)	0.293 (21)	0.167 (15)	0.362 (14)	0.639 (9)	0.203 (28)	0.733 (4)	0.235 (21)	0.616 (9)	0.254 (24)	0.746 (7)	0.218 (26)
Llama-4-Maverick	0.794 (4)	0.195 (22)	0.844 (2)	0.224 (24)	0.569 (8)	0.362 (14)	0.730 (6)	0.205 (28)	0.681 (6)	0.042 (27)	0.715 (5)	0.054 (30)	0.814 (6)	0.236 (25)
CommandR7B	0.729 (5)	0.254 (20)	0.747 (3)	0.254 (22)	0.179 (15)	0.362 (14)	0.614 (9)	0.187 (29)	0.671 (6)	0.185 (24)	0.457 (12)	0.202 (26)	0.451 (10)	0.188 (28)
Mistral-7B	0.620 (8)	0.296 (18)	0.598 (6)	0.301 (20)	0.170 (15)	0.362 (14)	0.534 (11)	0.229 (26)	0.600 (9)	0.230 (22)	0.448 (12)	0.222 (25)	0.487 (9)	0.242 (24)
AyaExpanse-8B	0.740 (5)	0.231 (21)	0.673 (5)	0.239 (23)	0.310 (11)	0.362 (14)	0.555 (10)	0.217 (27)	0.742 (4)	0.117 (26)	0.454 (12)	0.107 (29)	0.465 (10)	0.116 (30)

Table 21: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human2" as the gold standard. Metrics are ordered by category. Part 1 of 2.

Metric	en-it		en-ja		en-mas		en-ru		en-sr		en-uk		en-zh	
	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg	Sys	Seg
Baselines														
YiSi-1	–	–	0.734 (2)	0.516 (2)	–	–	0.585 (9)	0.518 (5)	0.840 (3)	0.651 (2)	0.702 (4)	0.521 (6)	0.658 (3)	0.503 (3)
chrF	–	–	0.734 (2)	0.513 (2)	–	–	0.583 (9)	0.515 (6)	0.817 (4)	0.638 (4)	0.736 (3)	0.526 (5)	0.677 (3)	0.506 (2)
spBLEU	–	–	0.711 (3)	0.508 (3)	–	–	0.632 (7)	0.521 (5)	0.847 (3)	0.642 (3)	0.705 (4)	0.517 (7)	0.657 (3)	0.503 (4)
BERTScore	–	–	0.729 (2)	0.508 (3)	–	–	0.596 (9)	0.514 (6)	0.821 (4)	0.642 (3)	0.655 (6)	0.507 (8)	0.633 (4)	0.496 (5)
BLEU	–	–	0.728 (2)	0.507 (3)	–	–	0.566 (10)	0.511 (6)	0.830 (4)	0.637 (4)	0.688 (5)	0.514 (7)	0.662 (3)	0.505 (2)
COMET22	–	–	0.576 (5)	0.498 (5)	–	–	0.571 (10)	0.516 (5)	0.835 (3)	0.654 (2)	0.698 (4)	0.539 (2)	0.584 (5)	0.503 (4)
sentinel-cand	0.665 (6)	0.507 (5)	0.479 (7)	0.468 (8)	0.443 (13)	0.673 (1)	0.570 (10)	0.496 (8)	0.781 (6)	0.603 (7)	0.633 (6)	0.511 (7)	0.483 (8)	0.477 (7)
COMETKiwi22	0.601 (7)	0.484 (8)	0.454 (8)	0.461 (9)	0.549 (10)	0.673 (1)	0.449 (11)	0.487 (10)	0.714 (8)	0.592 (8)	0.534 (8)	0.509 (8)	0.526 (6)	0.476 (7)
sentinel-src	0.422 (8)	0.172 (23)	0.524 (6)	0.157 (22)	0.547 (11)	0.673 (1)	0.556 (10)	0.142 (24)	0.455 (11)	0.155 (23)	0.563 (7)	0.155 (21)	0.551 (6)	0.161 (25)
Primary														
GEMBA-v2	0.830 (2)	0.540 (2)	0.773 (1)	0.526 (1)	0.558 (10)	0.673 (1)	0.835 (3)	0.559 (1)	0.891 (1)	0.644 (3)	0.836 (1)	0.541 (2)	0.744 (1)	0.494 (5)
TASER-No-Ref	0.807 (2)	0.486 (8)	0.781 (1)	0.454 (9)	0.696 (2)	0.673 (1)	0.852 (2)	0.503 (7)	0.890 (1)	0.618 (5)	0.841 (1)	0.477 (10)	0.721 (2)	0.451 (9)
rankedCOMET	0.663 (6)	0.502 (6)	0.573 (6)	0.499 (4)	0.615 (7)	0.673 (1)	0.583 (9)	0.516 (5)	0.829 (4)	0.655 (1)	0.704 (4)	0.539 (3)	0.594 (4)	0.504 (3)
MetricX-25	0.744 (3)	0.543 (2)	0.637 (4)	0.512 (3)	0.547 (11)	0.673 (1)	0.707 (6)	0.541 (3)	0.813 (4)	0.655 (1)	0.729 (4)	0.542 (2)	0.608 (4)	0.509 (2)
mr7_2_1	0.855 (1)	0.433 (11)	0.740 (2)	0.444 (11)	0.663 (4)	0.673 (1)	0.809 (3)	0.466 (12)	0.808 (5)	0.574 (9)	0.855 (1)	0.448 (11)	0.726 (2)	0.426 (11)
SEGALE-QE	0.743 (3)	0.536 (3)	0.531 (6)	0.476 (7)	0.636 (6)	0.673 (1)	0.596 (9)	0.514 (6)	0.666 (9)	0.561 (11)	0.714 (4)	0.528 (5)	0.589 (4)	0.486 (6)
Polycand-2	0.703 (5)	0.514 (4)	0.516 (7)	0.482 (6)	0.633 (6)	0.673 (1)	0.613 (9)	0.519 (5)	0.787 (6)	0.620 (5)	0.634 (6)	0.533 (4)	0.547 (6)	0.483 (6)
Q_Relative-MQM	0.825 (2)	0.347 (17)	0.731 (2)	0.384 (13)	0.626 (6)	0.673 (1)	0.783 (4)	0.366 (17)	0.852 (3)	0.498 (14)	0.805 (2)	0.352 (15)	0.759 (1)	0.379 (15)
EnsembleSlick	0.716 (5)	0.497 (6)	0.582 (5)	0.481 (7)	0.575 (9)	0.673 (1)	0.614 (9)	0.499 (8)	0.395 (11)	0.405 (18)	0.632 (6)	0.497 (9)	0.525 (7)	0.464 (8)
hw-tsc	0.656 (6)	0.494 (7)	0.473 (8)	0.465 (8)	0.588 (8)	0.673 (1)	0.461 (11)	0.499 (8)	0.744 (7)	0.600 (7)	0.585 (7)	0.513 (7)	0.521 (7)	0.480 (7)
UvA-MT	0.582 (7)	0.473 (9)	0.561 (6)	0.489 (6)	0.309 (14)	0.673 (1)	0.437 (11)	0.478 (11)	0.636 (9)	0.589 (8)	0.470 (9)	0.500 (9)	0.404 (9)	0.460 (8)
Roberta-LS	–	–	0.566 (6)	0.450 (10)	–	–	–	–	–	–	–	–	0.574 (5)	0.444 (10)
Secondary														
TASER-Ref	0.857 (1)	0.538 (2)	0.757 (1)	0.499 (4)	0.706 (2)	0.673 (1)	0.863 (2)	0.542 (3)	0.882 (1)	0.658 (1)	0.858 (1)	0.538 (3)	0.740 (1)	0.479 (7)
MetricX-25-Ref	–	–	0.648 (4)	0.517 (2)	–	–	0.726 (5)	0.549 (2)	0.812 (5)	0.656 (1)	0.756 (3)	0.558 (1)	0.641 (3)	0.515 (1)
baseCOMET	0.663 (6)	0.502 (6)	0.576 (6)	0.499 (4)	0.628 (6)	0.673 (1)	0.571 (10)	0.516 (5)	0.835 (4)	0.655 (1)	0.698 (4)	0.539 (2)	0.584 (5)	0.504 (3)
MetricX-25-QE	0.722 (4)	0.546 (1)	0.575 (6)	0.511 (3)	0.574 (9)	0.673 (1)	0.669 (7)	0.532 (4)	0.792 (6)	0.644 (3)	0.687 (5)	0.539 (2)	0.588 (5)	0.509 (2)
mr6	0.818 (2)	0.429 (11)	0.720 (2)	0.447 (10)	0.655 (5)	0.673 (1)	0.735 (4)	0.439 (14)	0.792 (5)	0.569 (10)	0.796 (2)	0.421 (12)	0.708 (2)	0.416 (12)
Q_MQM	0.826 (2)	0.352 (16)	0.729 (2)	0.394 (12)	0.505 (12)	0.673 (1)	0.785 (4)	0.370 (16)	0.858 (2)	0.505 (13)	0.798 (2)	0.358 (14)	0.766 (1)	0.390 (14)
Polyic-3	0.675 (6)	0.506 (5)	0.476 (8)	0.479 (7)	0.583 (8)	0.673 (1)	0.602 (9)	0.512 (6)	0.794 (5)	0.604 (7)	0.631 (6)	0.525 (6)	0.513 (7)	0.481 (6)
AutoLQA	0.774 (3)	0.426 (12)	0.691 (3)	0.398 (12)	0.577 (9)	0.673 (1)	0.733 (4)	0.426 (15)	0.815 (4)	0.466 (16)	0.760 (3)	0.392 (13)	0.586 (5)	0.333 (16)
Polycand-1	0.683 (6)	0.506 (5)	0.500 (7)	0.483 (6)	0.604 (7)	0.673 (1)	0.623 (8)	0.514 (6)	0.792 (6)	0.615 (6)	0.632 (6)	0.529 (5)	0.532 (6)	0.480 (7)
CollabPlus	0.702 (5)	0.514 (4)	0.559 (6)	0.487 (6)	0.531 (11)	0.673 (1)	0.602 (9)	0.505 (7)	0.412 (11)	0.396 (19)	0.654 (6)	0.509 (8)	0.518 (7)	0.479 (7)
CollabSlick	0.726 (4)	0.513 (4)	0.585 (5)	0.488 (6)	0.566 (10)	0.673 (1)	0.622 (8)	0.508 (7)	0.392 (11)	0.400 (19)	0.638 (6)	0.504 (8)	0.538 (6)	0.475 (7)
hw-tsc-max	0.594 (7)	0.483 (8)	0.466 (8)	0.458 (9)	0.580 (9)	0.673 (1)	0.431 (12)	0.492 (9)	0.683 (9)	0.571 (9)	0.552 (8)	0.502 (9)	0.512 (7)	0.478 (7)
hw-tsc-base	0.594 (7)	0.483 (8)	0.466 (8)	0.458 (9)	0.580 (9)	0.673 (1)	0.431 (12)	0.492 (9)	0.683 (9)	0.571 (9)	0.552 (8)	0.502 (9)	0.512 (7)	0.478 (7)
long-context	–	–	0.566 (6)	0.485 (6)	–	–	–	–	–	–	–	–	0.550 (6)	0.438 (10)
roberta-multi	–	–	0.523 (7)	0.449 (10)	–	–	–	–	–	–	–	–	0.550 (6)	0.453 (9)
LLM-as-a-judge														
GPT-4_1	0.826 (2)	0.456 (10)	0.758 (1)	0.450 (10)	0.695 (3)	0.673 (1)	0.863 (2)	0.457 (13)	0.871 (2)	0.550 (12)	0.842 (1)	0.446 (11)	0.733 (2)	0.407 (13)
CommandA	0.829 (2)	0.404 (13)	0.773 (1)	0.341 (14)	0.694 (3)	0.673 (1)	0.891 (1)	0.366 (17)	0.823 (4)	0.477 (15)	0.843 (1)	0.360 (14)	0.717 (2)	0.322 (18)
Claude-4	0.842 (1)	0.300 (18)	0.745 (2)	0.301 (16)	0.702 (2)	0.673 (1)	0.830 (3)	0.285 (20)	0.864 (2)	0.457 (17)	0.830 (2)	0.264 (18)	0.737 (1)	0.268 (20)
DeepSeek-V3	0.843 (1)	0.349 (16)	0.755 (1)	0.295 (16)	0.719 (1)	0.673 (1)	0.838 (3)	0.334 (19)	0.832 (4)	0.496 (14)	0.855 (1)	0.318 (17)	0.717 (2)	0.330 (17)
Qwen3-235B	0.843 (1)	0.386 (14)	0.730 (2)	0.333 (15)	0.666 (4)	0.673 (1)	0.863 (2)	0.357 (18)	0.825 (4)	0.459 (17)	0.856 (1)	0.337 (16)	0.744 (1)	0.338 (16)
Qwen2_5-7B	0.816 (2)	0.378 (15)	0.725 (2)	0.294 (17)	0.662 (4)	0.673 (1)	0.776 (4)	0.361 (17)	0.749 (7)	0.396 (19)	0.679 (5)	0.348 (15)	0.716 (2)	0.310 (19)
AyaExpanse-32B	0.824 (2)	0.285 (19)	0.697 (3)	0.230 (21)	0.661 (4)	0.673 (1)	0.774 (4)	0.226 (23)	0.702 (8)	0.284 (20)	0.776 (3)	0.225 (19)	0.672 (3)	0.215 (23)
Llama-3_1-8B	0.766 (3)	0.242 (21)	0.666 (3)	0.236 (20)	0.613 (7)	0.673 (1)	0.681 (6)	0.243 (21)	0.759 (7)	0.268 (21)	0.672 (5)	0.217 (20)	0.677 (3)	0.258 (21)
Llama-4-Maverick	0.759 (3)	0.116 (24)	0.700 (3)	0.101 (24)	0.660 (4)	0.673 (1)	0.689 (6)	0.071 (26)	0.733 (7)	0.175 (22)	0.619 (6)	0.070 (23)	0.691 (3)	0.089 (26)
CommandR7B	0.765 (3)	0.256 (20)	0.708 (3)	0.266 (18)	0.677 (4)	0.673 (1)	0.550 (10)	0.238 (22)	0.355 (12)	0.150 (24)	0.487 (8)	0.228 (19)	0.593 (4)	0.264 (20)
Mistral-7B	0.615 (7)	0.252 (20)	0.514 (7)	0.254 (19)	0.646 (5)	0.673 (1)	0.535 (10)	0.245 (21)	0.599 (10)	0.287 (20)	0.545 (8)	0.229 (19)	0.534 (6)	0.226 (22)
AyaExpanse-8B	0.693 (5)	0.180 (22)	0.715 (2)	0.140 (23)	0.656 (4)	0.673 (1)	0.592 (9)	0.130 (25)	0.404 (11)	0.157 (23)	0.554 (7)	0.130 (22)	0.660 (3)	0.167 (24)

Table 22: System- and segment-level correlations per language pair for Task 1, with rankings shown in parentheses, computed against "human2" as the gold standard. Metrics are ordered by category. Part 2 of 2.

C Task 1 Score Difference Interpretation Additional Figures

Figures 9-11 show the (log) p-value of one-sided paired t-test on the human scores against the score difference of each auto-rater for each system pair. Figures 12-13 show the (log) p-value of significance test with bootstrap resampling on the auto-rater scores against the score difference of that auto-rater for each system pair in each language pair.

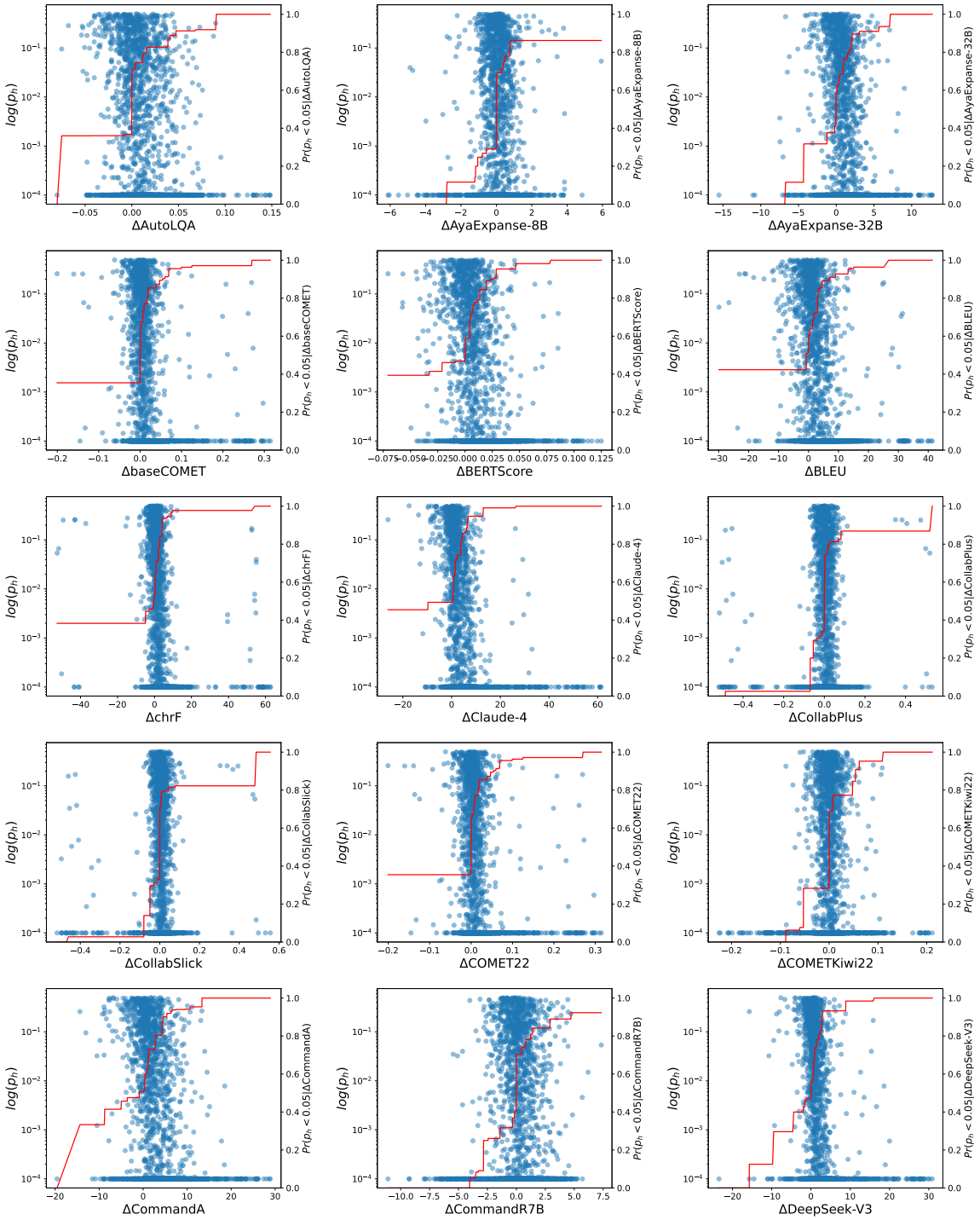


Figure 9: Log p-value of one-sided paired t-test on MQM scores (p_h) against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \mid \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001. (Part 1/3)

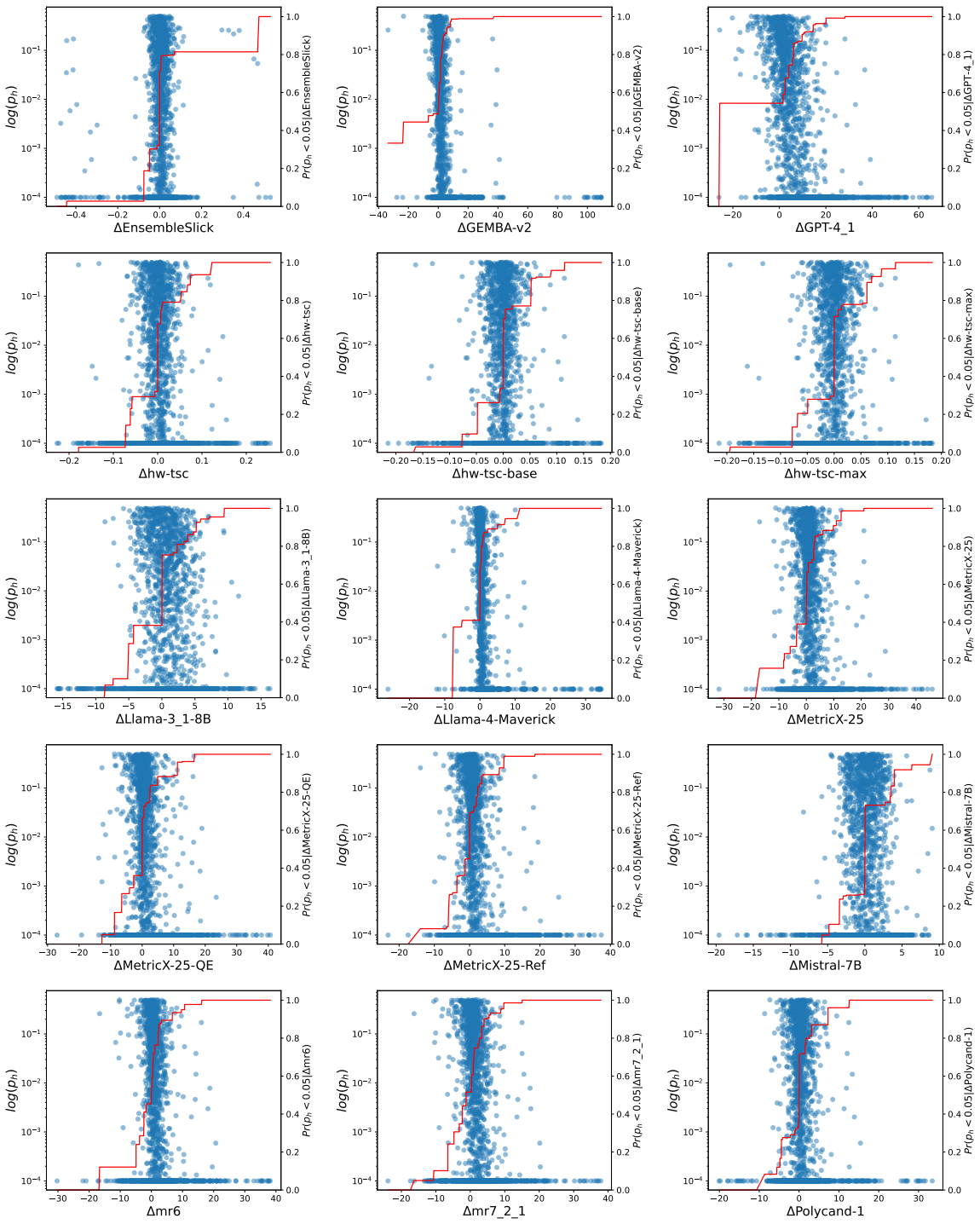


Figure 10: Log p-value of one-sided paired t-test on MQM scores (p_h) against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \mid \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001.(Part 2/3)

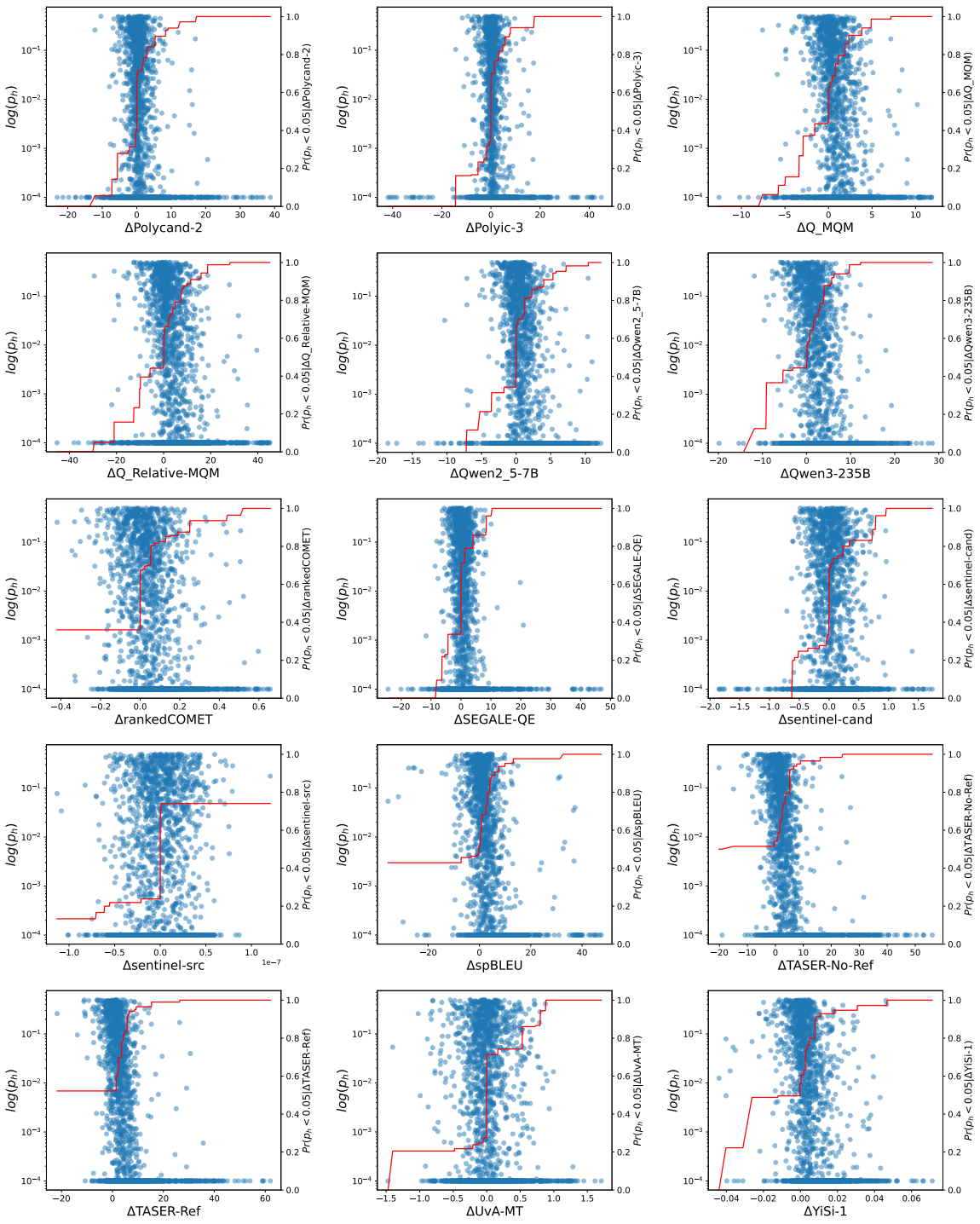


Figure 11: Log p-value of one-sided paired t-test on MQM scores (p_h) against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_h < 0.05 \mid \Delta m)$. Note: for readability, values of p_h are rounded up to 0.0001 when they are less than 0.0001.(Part 3/3)

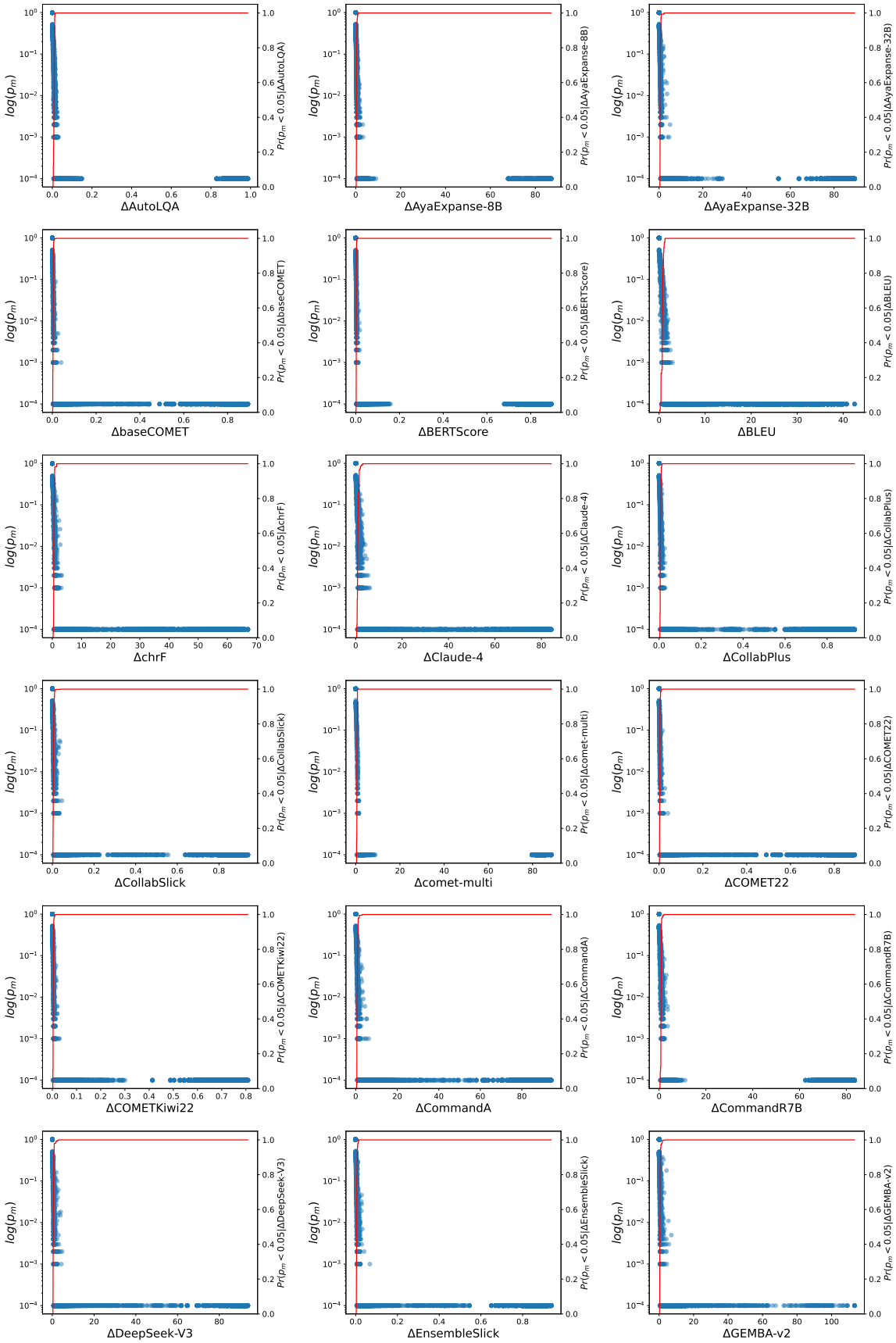


Figure 12: Log p-value of significance test with bootstrap resampling (p_m) on system-level against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \mid \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001. (Part 1/3)

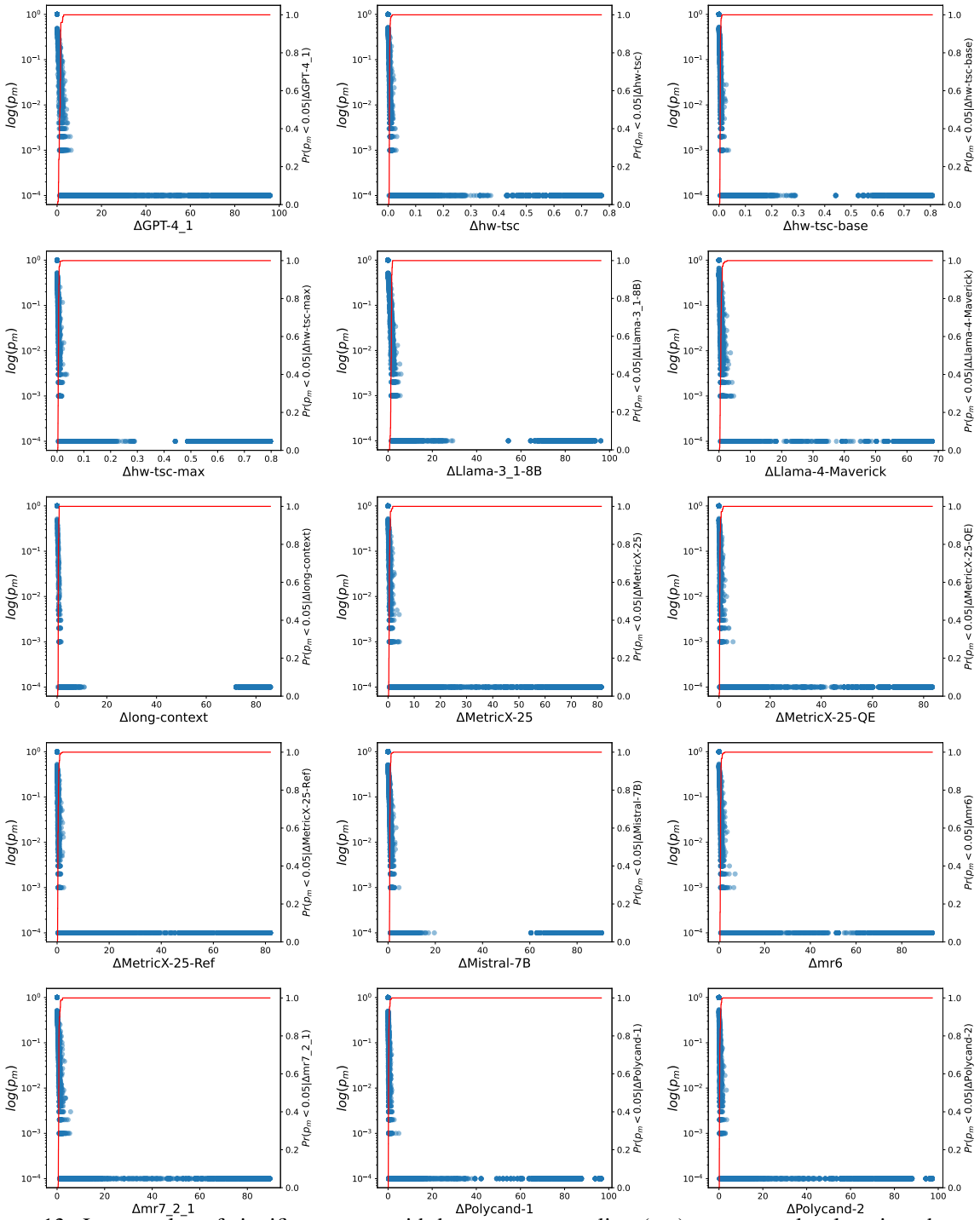


Figure 13: Log p-value of significance test with bootstrap resampling (p_m) on system-level against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \mid \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001. (Part 2/3)

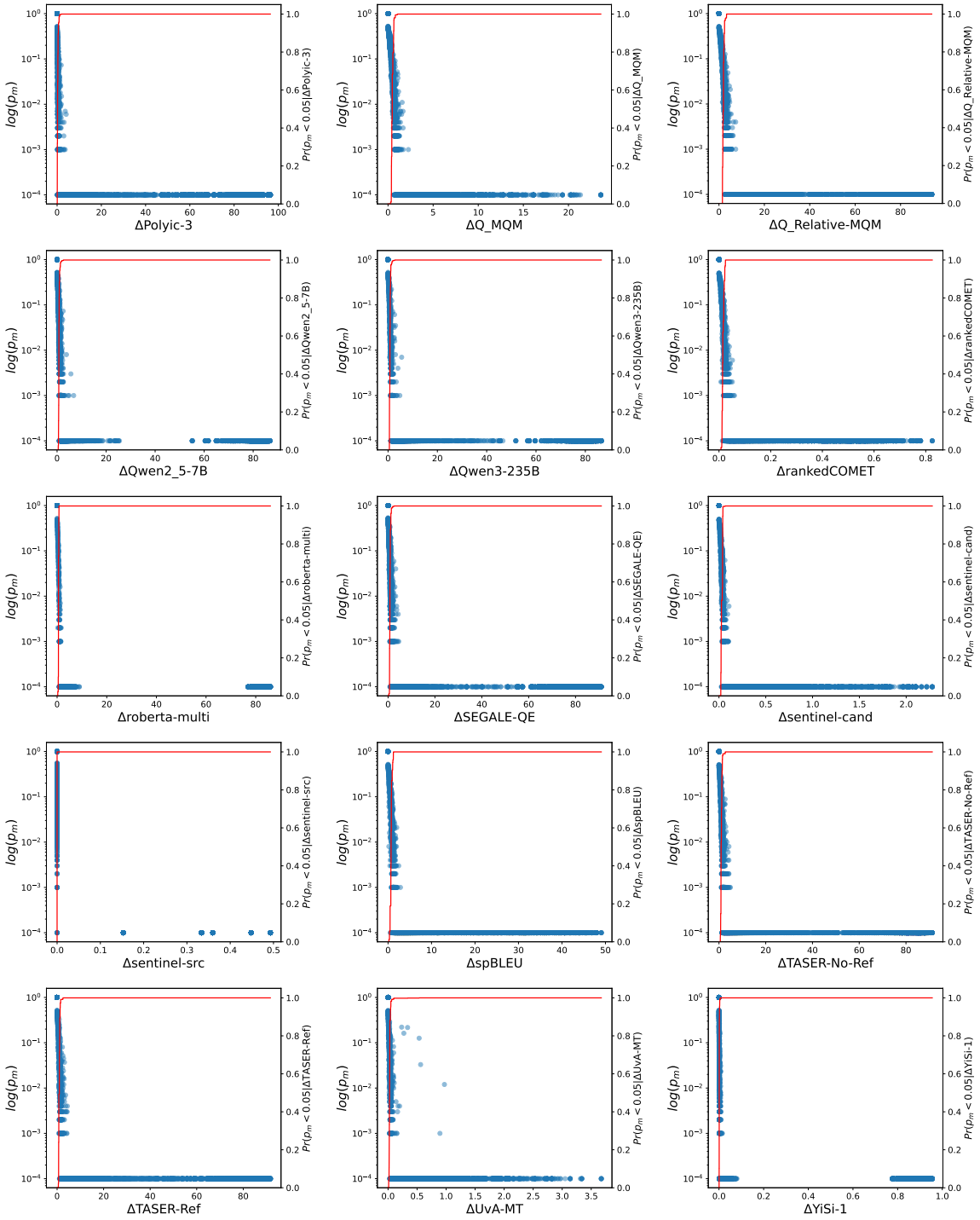


Figure 14: Log p-value of significance test with bootstrap resampling (p_m) on system-level against the score difference of each auto-rater for each system pair. The red line is the isotonic regression fit to all data points, representing $Pr(p_m < 0.05 \mid \Delta m)$. Note: for readability, values of p_m are rounded up to 0.0001 when they are less than 0.0001. (Part 3/3)

D Task 2 Additional Details

Algorithm 1 Character-Level Error Span F1 Score

function GET_CHAR_F1($L_{hyp}, E_{\text{gold}}, E_{\text{pred}}, \rho$)
▷ L_{hyp}: List of hypothesis lengths
▷ E_{gold}: List of lists of gold error dicts
▷ E_{pred}: List of lists of predicted error dicts
▷ ρ: Partial credit factor (e.g., 0.5)
    $tp \leftarrow 0$
    $\text{total\_gold} \leftarrow 0$
    $\text{total\_pred} \leftarrow 0$
    for $i \in 0 \ldots \text{LENGTH}(L_{hyp}) - 1$ **do**
        $H_{\text{len}} \leftarrow L_{hyp}[i]$
        $G_{\text{maj}}, G_{\text{min}} \leftarrow \text{GET\_COUNTS}(E_{\text{gold}}[i], H_{\text{len}})$
        $P_{\text{maj}}, P_{\text{min}} \leftarrow \text{GET\_COUNTS}(E_{\text{pred}}[i], H_{\text{len}})$
        $\text{total\_gold} \leftarrow \text{total\_gold} + \sum G_{\text{maj}} + \sum G_{\text{min}}$
        $\text{total\_pred} \leftarrow \text{total\_pred} + \sum P_{\text{maj}} + \sum P_{\text{min}}$
        for $j \leftarrow 0$ To $H_{\text{len}} - 1$ **do**
            $c_{\text{g\_maj}} \leftarrow G_{\text{maj}}[j]$
            $c_{\text{p\_maj}} \leftarrow P_{\text{maj}}[j]$
            $c_{\text{g\_min}} \leftarrow G_{\text{min}}[j]$
            $c_{\text{p\_min}} \leftarrow P_{\text{min}}[j]$
▷ Full credit for same severity match at index j
            $tp \leftarrow tp + \min(c_{\text{g\_maj}}, c_{\text{p\_maj}})$
            $tp \leftarrow tp + \min(c_{\text{g\_min}}, c_{\text{p\_min}})$
▷ Partial credit for cross-severity match at index j
            $g_{\text{unmatched}} \leftarrow \max(0, c_{\text{g\_maj}} - c_{\text{p\_maj}}) + \max(0, c_{\text{g\_min}} - c_{\text{p\_min}})$
            $p_{\text{unmatched}} \leftarrow \max(0, c_{\text{p\_maj}} - c_{\text{g\_maj}}) + \max(0, c_{\text{p\_min}} - c_{\text{g\_min}})$
            $tp \leftarrow tp + \min(g_{\text{unmatched}}, p_{\text{unmatched}}) \times \rho$
        end for
    end for
    $P, R, F1 \leftarrow \text{PREC\_REC\_F1}(tp, \text{total\_gold}, \text{total\_pred})$
    return $P, R, F1$
end function

Figure 15: F1 Score by Error Category.

E Full Results for Task 2

We report complete results for all submissions (primary and secondary) in Tables 23 (micro-F1) and 24 (macro-F1) respectively. We also show macro-F1 scores broken down by error category and error ratio in Figure 15 and Figure 16 respectively.

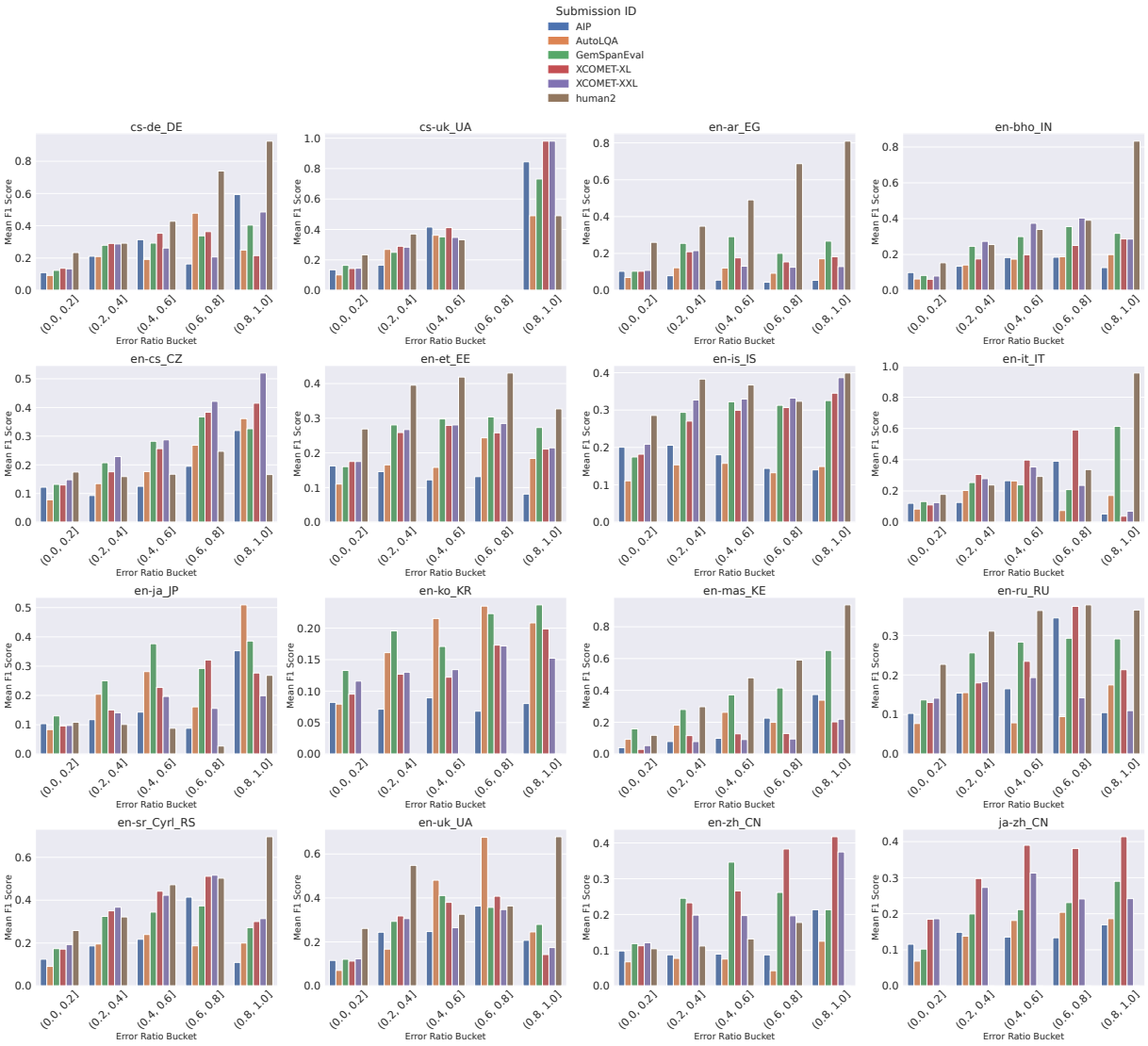


Figure 16: F1 Score by Error Ratio.

Language Pair	Baselines						Primary Submissions									Secondary Submissions									Human2		
	XCOMET-XL			XCOMET-XXL			AutoLQA			AIP			GemSpanEval			AutoLQA			AIP			GemSpanEval					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CS-DE_DE	24.55	5.15	8.52	25.17	5.22	8.65	17.71	4.02	6.56	11.94	20.37	15.06	28.89	6.02	9.96	15.54	4.58	7.08	14.76	20.31	17.09	31.33	6.66	10.99	30.46	41.08	34.98
CS-UK_UA	25.02	4.00	6.90	28.21	3.56	6.32	20.73	1.93	3.54	13.60	10.16	11.63	35.94	3.58	6.52	17.02	1.99	3.56	18.71	12.03	14.65	37.31	3.96	7.15	27.67	28.95	28.30
EN-AR_EG	11.57	19.53	14.54	8.48	17.92	11.51	11.45	22.95	15.28	2.51	30.41	4.63	19.23	22.18	20.60	10.28	22.65	14.14	2.95	31.76	5.40	21.16	23.50	22.27	79.61	76.37	77.96
EN-BHO_IN	22.87	3.47	6.02	33.46	3.40	6.17	15.55	3.48	5.69	9.38	8.19	8.74	28.40	3.68	6.52	14.91	3.45	5.60	5.54	6.70	6.07	24.57	3.70	6.43	61.31	54.03	57.44
EN-CS_CZ	16.06	6.02	8.76	22.67	6.87	10.55	14.22	4.12	6.39	7.27	15.30	9.85	24.20	6.38	10.10	15.40	3.98	6.32	9.45	15.68	11.79	25.93	6.97	10.99	14.40	24.86	18.24
EN-ET_EE	14.93	16.66	15.75	16.56	17.07	16.81	14.84	11.82	13.16	5.71	20.34	8.92	23.09	12.28	16.04	11.41	10.86	11.13	6.75	22.86	10.42	22.97	12.54	16.23	33.31	32.87	33.09
EN-IS_IS	22.41	27.72	24.78	29.10	28.30	28.70	8.85	19.68	12.21	9.15	35.69	14.57	25.90	19.58	22.30	6.65	17.82	9.69	9.70	34.50	15.14	25.90	20.12	22.64	36.44	40.10	38.18
EN-IT_IT	30.60	4.16	7.33	23.40	5.40	8.77	17.89	2.55	4.47	10.45	13.70	11.86	33.71	5.47	9.41	18.66	2.69	4.71	11.92	15.02	13.29	33.98	5.53	9.51	30.52	30.62	30.57
EN-JA_JP	13.97	3.67	5.81	14.85	3.69	5.92	22.70	2.30	4.18	8.88	11.72	10.10	28.47	3.32	5.94	23.88	2.71	4.86	9.61	10.60	10.08	28.64	3.26	5.85	10.61	13.93	12.04
EN-KO_KR	8.74	14.64	10.95	9.96	17.09	12.58	20.23	7.26	10.69	4.81	25.89	8.12	17.65	10.54	13.20	16.50	6.91	9.75	5.42	28.15	9.08	20.78	11.17	14.53	-	-	-
EN-MAS_KE	10.23	34.84	15.81	11.35	36.31	17.29	15.14	38.95	21.80	27.94	28.65	28.29	35.03	35.67	35.35	14.92	37.87	21.41	4.43	38.08	7.93	35.14	35.72	35.43	94.73	92.14	93.41
EN-RU_RU	16.70	8.59	11.34	17.95	8.48	11.52	13.77	3.84	6.01	8.74	16.77	11.49	28.28	6.49	10.55	12.32	4.00	6.03	9.72	17.06	12.38	29.29	6.39	10.49	25.28	27.56	26.37
EN-SR_CYRL	21.18	21.59	21.38	24.17	21.07	22.52	13.61	15.16	14.35	6.81	27.11	10.88	21.66	15.67	18.18	11.76	15.19	13.26	7.56	26.42	11.75	21.23	17.15	18.97	61.67	58.32	59.95
EN-UK_UA	21.84	2.98	5.25	27.31	3.20	5.73	15.48	1.32	2.43	12.63	6.98	8.99	37.01	2.19	4.13	14.13	1.22	2.24	14.37	6.89	9.32	35.19	2.25	4.22	34.76	39.28	36.88
EN-ZH_CN	22.62	3.80	6.50	19.45	4.14	6.83	13.08	2.92	4.78	7.72	10.43	8.87	30.02	3.37	6.07	11.61	3.19	5.01	9.00	10.40	9.65	30.30	3.47	6.22	11.84	12.82	12.31
JA-ZH_CN	26.89	21.80	24.08	24.07	20.04	21.87	14.83	13.58	14.18	8.47	41.64	14.08	25.35	17.46	20.68	13.11	13.00	13.06	8.50	39.40	13.98	27.67	17.23	21.23	-	-	-
Average	19.39	12.41	12.11	21.01	12.61	12.61	15.63	9.74	9.11	9.75	20.21	11.63	27.68	10.87	13.47	14.26	9.51	8.61	9.27	20.99	11.13	28.21	11.23	13.95	47.04†	48.31†	47.48†

Table 23: Task 2 micro-F1 (%) by language pair for all auto-raters. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

Language Pair	Baselines						Primary Submissions									Secondary Submissions									Human2		
	XCOMET-XL			XCOMET-XXL			AutoLQA			AIP			GemSpanEval			AutoLQA			AIP			GemSpanEval					
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
CS-DE_DE	70.91	16.76	13.07	70.82	21.97	16.22	67.13	20.05	13.91	64.51	55.74	36.45	69.95	23.21	17.08	65.66	24.67	16.63	66.03	46.16	31.22	71.35	24.46	19.14	70.56	77.94	64.46
CS-UK_UA	75.93	20.87	17.49	76.62	24.02	19.37	74.50	16.09	11.44	72.11	49.75	37.74	78.98	17.35	15.67	73.35	17.57	11.99	73.87	40.87	32.48	79.37	18.09	16.33	75.86	80.40	67.55
EN-AR_EG	60.85	25.16	10.54	59.32	27.77	10.07	59.48	31.81	11.53	55.38	49.47	18.86	64.28	28.15	12.24	59.01	32.19	10.12	55.70	42.66	14.04	65.07	27.87	12.89	84.95	82.04	79.33
EN-BHO_IN	83.11	27.92	20.00	86.97	13.37	9.88	79.92	10.80	7.14	77.53	32.68	22.44	83.93	8.92	7.01	79.93	8.65	5.46	77.66	21.92	13.40	82.44	10.13	7.71	82.96	83.15	78.86
EN-CS_CZ	71.64	14.84	10.79	73.77	13.70	10.93	70.09	27.18	17.36	68.01	47.65	31.78	73.59	16.22	12.68	70.61	19.95	13.14	69.36	37.75	25.72	74.33	17.54	14.28	70.92	74.52	60.70
EN-ET_EE	66.82	20.71	11.62	67.00	23.50	13.72	65.39	23.96	12.97	62.35	47.05	24.89	70.18	15.65	10.83	64.16	23.35	11.43	63.41	35.98	18.18	70.65	15.69	11.17	76.91	71.50	64.77
EN-IS_IS	60.19	26.54	15.87	63.13	26.79	18.50	52.49	26.35	10.01	53.37	47.42	21.83	62.36	22.23	14.95	51.04	26.97	9.59	54.43	39.01	17.22	63.01	22.05	15.24	68.79	72.75	62.51
EN-IT_IT	71.04	7.13	7.65	66.96	13.38	11.41	64.67	17.00	11.51	60.63	50.35	32.03	68.60	13.82	11.92	64.81	17.29	11.93	61.41	45.63	29.34	68.48	13.89	11.91	65.00	63.88	52.23
EN-JA_JP	77.58	13.98	10.67	77.54	20.40	16.15	79.76	14.14	11.50	76.13	58.14	44.81	80.78	10.04	8.33	79.75	12.77	10.39	77.02	48.80	37.91	80.88	13.94	12.15	76.78	78.98	64.29
EN-KO_KR	45.43	28.05	12.27	45.99	27.36	14.32	50.64	28.32	14.05	42.77	60.19	26.44	51.27	19.28	11.77	48.44	33.92	15.36	43.31	56.78	24.98	53.33	20.63	13.27	-	-	-
EN-MAS_KE	69.29	77.53	49.07	70.33	77.34	49.19	73.73	48.53	27.42	75.57	57.75	36.13	85.55	39.78	31.45	73.76	47.13	26.88	67.09	44.18	15.59	85.56	39.88	31.59	97.13	96.40	96.33
EN-RU_RU	64.66	17.65	13.19	64.86	18.70	13.80	62.52	30.86	19.09	60.07	50.32	30.50	68.39	12.33	10.77	61.55	29.45	18.06	60.85	45.19	27.29	69.13	12.75	11.20	67.48	70.32	58.49
EN-SR_CYRL_RS	65.00	21.77	14.08	66.36	23.09	15.19	60.11	36.45	18.72	57.26	47.01	23.76	66.38	18.92	12.05	59.51	34.75	16.94	58.13	34.44	15.59	66.33	18.95	12.22	72.39	73.76	64.19
EN-UK_UA	79.47	11.05	10.17	80.71	11.26	10.79	77.33	22.65	16.97	76.73	43.18	33.69	82.64	6.07	6.55	77.15	13.71	10.07	77.73	34.16	27.20	82.42	6.19	6.37	80.58	83.82	72.02
EN-ZH_CN	77.07	7.00	6.55	75.95	10.11	8.65	73.00	45.64	32.37	71.78	52.18	38.14	78.53	8.33	7.50	72.27	45.62	32.29	72.51	46.02	33.65	78.76	8.98	8.01	72.96	74.32	59.82
JA-ZH_CN	49.92	28.45	20.17	48.18	31.36	19.78	37.71	50.95	18.90	34.94	66.17	25.83	42.09	56.08	25.74	36.42	49.17	18.48	36.03	61.77	25.03	43.08	56.51	26.76	-	-	-
Average	68.06	22.84	15.20	68.41	24.01	16.12	65.53	28.17	15.93	63.07	50.94	30.33	70.47	19.77	13.53	64.84	27.32	14.92	63.41	42.58	24.30	70.89	20.47	14.39	78.96†	80.24†	71.60†

Table 24: Task 2 macro-F1 (%) by language pair for all auto-raters. †: average is computed over all but JA-ZH_CN and EN-KO_KR.

F Prompts Used for Task 3

```
Translate the following from source_lang to target_lang.  Include only the
translation (without the <>) and nothing else.
>source_text<
```

Table 25: Prompt for Quality-Informed Segment-Level Error Correction task with translating from scratch.

```
Post-edit the following translation from source_lang to target_lang:
>original_translation< given these errors error_spans.  Include only the translation
(without the <>) and nothing else.
>source_text<
```

Table 26: Prompt for Quality-Informed Segment-Level Error Correction task with post-editing existing translation.

Findings of the WMT 2025 Shared Task on Model Compression: Early Insights on Compressing LLMs for Machine Translation

Marco Gaido FBK
Thamme Gowda Microsoft
Roman Grundkiewicz Microsoft
Matteo Negri FBK

{mgaido,negri}@fbk.eu, {thammegowda,rogrundk}@microsoft.com

Abstract

We present the results of the first edition of the Model Compression shared task, organized as part of the 10th Conference on Machine Translation (WMT25). The task challenged participants to compress Large Language Models (LLMs) toward enabling practical deployment in resource-constrained scenarios, while minimizing loss in translation performance. In this edition, participants could choose to compete in either a constrained track, which required compressing a specific model (Aya Expanse 8B) evaluated on a limited set of language pairs (Czech→German, Japanese→Chinese, and English→Arabic), or an unconstrained track, which placed no restrictions on the model and allowed submissions for any of the 15 language directions covered by the General MT task. We received 12 submissions from 3 teams, all in the constrained track. They proposed different compression solutions and covered various language combinations. Evaluation was conducted separately for each language, measuring translation quality using COMET and MetricX, model size, and inference speed on an Nvidia A100 GPU.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional performance across a wide range of tasks. However, efforts to enhance their capabilities, by expanding language coverage, integrating multimodal data, and improving task generalization, have led to a dramatic increase in both model size and computational demands (Zhu et al., 2024). This rapid growth poses significant challenges for real-world deployment, particularly in resource-constrained environments such as mobile devices, embedded systems, and edge computing platforms, where low-latency, on-device processing is often required. Compressing foundation models is therefore more than a technical pursuit: it is a strategic priority with implications for global accessibility[1] and the sustainability of computational and environmental costs. Striking the right balance between performance, compactness, and efficiency is thus essential to make LLMs truly ubiquitous and beneficial for everyone, regardless of location or access to high-end infrastructure.

It is with this long-term goal in mind that, following the analogous task for speech translation in the IWSLT 2025 campaign (Abdulmumin et al., 2025), the new Model Compression shared task was introduced at WMT 2025.[2] This initiative follows three editions of the shared task at the Workshop on Machine Translation and Generation (Birch et al., 2018; Hayashi et al., 2019; Heafield et al., 2020) and two editions of the shared task on Efficient Translation (Heafield et al., 2021, 2022). It revives earlier focus on the efficiency of machine translation, while updating it to reflect the current AI landscape with the rise of general-purpose LLMs.

In this context, our aim is to provide a timely evaluation of compression techniques for general-purpose LLMs within the specific task of machine translation. This setting offers a valuable opportunity to explore key research questions, such as:

- *To what extent can the over-parameterization of LLMs—originally pursued to enable generalization, robustness, task flexibility, and broad language coverage—be reduced in favor of compactness and efficiency, while preserving MT quality?*
- *How do different compression techniques, with varying degrees of aggressiveness, impact translation quality in such settings?*

[1] In the U.S. around 15% of adults rely exclusively on mobile devices to access the internet (https://www.pewresearch.org/internet/fact-sheet/mobile/), and it is even more pronounced in developing regions (https://www.eib.org/en/essays/african-digital-infrastructure).

[2] https://www2.statmt.org/wmt25/

Proceedings of the Tenth Conference on Machine Translation, pages 484–494
November 8-9, 2025

2 Task Description

The goal of the Model Compression task is to reduce the size of a general-purpose LLM while preserving a strong balance between compactness and MT performance. This section provides a brief overview of how the first round of the task was structured, focusing on the proposed tracks, data conditions, and evaluation methodology.

2.1 Tracks

Participants could choose between two tracks: constrained and unconstrained.

The **constrained** track was designed to ensure a level playing field by establishing uniform conditions across all participants, allowing for directly comparable results. It focused on the compression of a specific model in a fixed language setting. The model selected for this purpose was Aya Expanse 8B,[3] chosen for its permissive license (CC-BY-NC 4.0) and its favorable trade-off between the size (8 billion parameters; approximately 16 GB in FP16 precision) and performance.

In the constrained settings, we measured performance across three language pairs: Czech→German, Japanese→Chinese, English→Arabic. These pairs were selected to provide a sufficiently diverse coverage of language families and scripts. Submissions were allowed for any of these directions. Any model compression technique e.g., pruning (Frankle and Carbin, 2019; Frankle et al., 2020), quantization (Devlin, 2017), or distillation (Kim and Rush, 2016), was permitted, provided that the final compressed model remained closely derived from Aya Expanse 8B. For instance, in the case of distillation, student models had to be obtained through compression of Aya Expanse 8B (e.g., by pruning or quantizing it) to qualify for the constrained track. Otherwise, we would consider such systems as unconstrained.

The **unconstrained** track provided participants with complete freedom to compress any model of their choice and apply it to any of the 15 language directions covered by the WMT25 General MT task (GenMT) (Kocmi et al., 2025). As in the constrained track, separate rankings were planned for each language direction.

[3] https://huggingface.co/CohereLabs/aya-expanse-8b

2.2 Data

Data usage policies were aligned with those of the GenMT task. Participants were therefore allowed to calibrate and fine-tune their compressed models using the publicly available datasets released for this year's round,[4] as well as test sets from previous WMT editions.

2.3 Evaluation

Submissions were evaluated[5] along three key dimensions:

- **Translation quality** measured using the same automatic metrics employed in the GenMT task;
- **Model size** as disk space footprint;
- **Inference speed** as the average number of output tokens produced per second when processing the test set.

All three were considered both *independently* and *jointly*. We report Pareto frontier rankings to visualize system differences through quality–size, quality–speed and size–speed plots. Since we received multiple submissions only for Czech→German, this type of visualization was only feasible for that language direction.

To ensure a fair and informative evaluation, we create a homogeneous hardware environment for running the submitted systems. We used machines with a single Nvidia A100 GPU having 80GB of VRAM, AMD EPYC CPU with 96 cores, and 866GB RAM.

2.4 Submission

Participating teams were asked to provide a link to a Docker image containing all necessary software and model files for translation, along with basic information about the maximum batch size supported by their model(s) under the specified hardware configuration. Upon request, we also offered storage space to teams who needed it or preferred to upload their models externally to their institutional infrastructure.

3 Participants

Three teams submitted systems to the task, as summarized in Table 1. The organizers also included baseline systems. Below, we provide a brief

[4] https://www2.statmt.org/wmt25/mtdata/

[5] Scripts used for evaluation are available at: https://github.com/thammegowda/wmt25-model-compression

Institution	Submission	Track	No. Sub.	Languages
Stevens Institute of Technology, Rice University, Lambda Inc.	AyaQ	Constr.	1	cs-de
Stevens Institute of Technology, Rice University, Lambda Inc.	LeanAya	Constr.	1	cs-de
Trinity College Dublin (Moslem et al., 2025)	TCD-Kreasof	Constr.	3	cs-de
Vicomtech (Ponce et al., 2025b)	Vicomtech	Constr.	7	cs-de, jp-zh, en-ar
Organizers (compressed baseline model)	BitsAndBytes	Constr.	4	*as baseline*

Table 1: Participants in the WMT 2025 Model Compression shared task with the number of submitted system variants and declared language support.

Submission	Description
base	Base Aya Expanse 8B in 16bit
bnb-8bit	8bit integer
bnb-4bit-fp4	4bit FP4
bnb-4bit-nf4	4bit NF4
bnb-4bit-nf4-2q	4bit NF4, double-quantization

Table 2: Baseline and BitsAndBytes (Dettmers et al., 2022, 2023) systems submitted by the organizers.

overview of the proposed approaches, all developed within the constrained track.

AyaQ[6] This participation employs GPTQ 4-bit quantization (Frantar et al., 2023) with a group size of 32 to enable efficient and scalable LLM inference. The WMT dataset is used as calibration data to guide the quantization process, ensuring the compressed model retains high accuracy on language understanding and generation tasks. The quantized models are integrated through the LLM Compressor framework (AI and vLLM Project, 2024), which streamlines conversion and metadata management. The setup is fully compatible with vLLM (Kwon et al., 2023), a high-throughput inference engine optimized for GPU deployment, enabling fast and memory-efficient execution with minimal performance loss. This approach demonstrates how structured quantization, targeted calibration, and system-level integration can enable practical, production-ready LLM deployment.

LeanAya This participation is based on LeanQuant (Loss-Error-Aware Network Quantization, (Zhang and Shrivastava, 2025)), an accurate, versatile, and scalable quantization method. Existing iterative loss-error-based quantization techniques typically rely on min-max affine grids, which often degrade model quality due to outliers in the inverse Hessian diagonals. LeanQuant overcomes this limitation by learning loss-error-aware quantization grids instead of using fixed, non-adaptive ones. This approach not only improves accuracy but also supports a wider range of quantization schemes, including both affine and non-uniform, enhancing compatibility across diverse deployment frameworks.

TCD-Kreasof (Moslem et al., 2025) This participation employs iterative layer pruning to incrementally identify and remove layers that contribute least to translation quality, one at a time. Layer importance is assessed by measuring translation performance with each layer individually removed. After pruning the least critical layer, the evaluation is repeated on the remaining ones until the target pruning level is reached. The resulting pruned model was then fine-tuned on 100k sentences from the News Commentary dataset. This process produced three submissions: the primary one is a 24-layer model with 6.28B parameters, while the two contrastive submissions are 20-layer and 16-layer models, with 5.41B and 4.54B parameters, respectively.

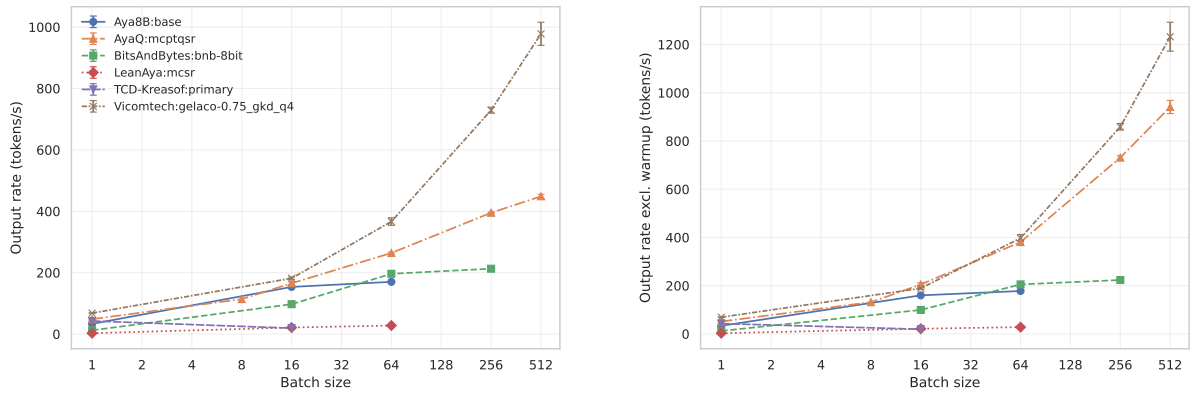
Vicomtech (Ponce et al., 2025b) This participation employs GeLaCo (Ponce et al., 2025a), an evolutionary approach to LLM compression based on layer merging operations. Models are compressed at three ratios (0.25, 0.50, and 0.75), representing the proportion of original layers collapsed through differential weight merging. To recover performance after compression, over 3 million translation instructions (1 million per language) from a subset of WMT25 translation data are used. For the 0.25 and 0.50 compression levels, models are fine-tuned on this data, while the 0.75 model is trained using General Knowledge Distillation (GKD (Tan et al., 2023)). Additionally, post-training quantization is applied using the `bitsandbytes` library[7] to further reduce model size to 8-bit and 4-bit precision. The primary submission (`gelaco-0.75_gkd_q4`)

[6] We did not receive system description papers for AyaQ and LeanAya submissions.

[7] https://github.com/bitsandbytes-foundation/bitsandbytes

		English→Arabic		Japanese→Chinese		Czech→German		
Submission	**System**	**COMET↑**	**MetricX↓**	**COMET↑**	**MetricX↓**	**COMET↑**	**MetricX↓**	**#Halluc.**
Baseline	base	25.4	8.20	44.5	6.44	55.3	5.08	83
BitsAndBytes	**bnb-8bit**	25.2	8.33	44.4	6.49	55.6	5.08	86
	bnb-4bit-fp4	24.4	8.26	43.6	6.54	54.5	5.24	153
	bnb-4bit-nf4	25.4	8.25	44.6	6.43	55.7	5.15	103
	bnb-4bit-nf4-2q	25.4	8.25	44.6	6.41	55.5	5.15	106
AyaQ	**mcptqsr**	–	–	–	–	40.3	8.70	200
LeanAya	**mcsr**	–	–	–	–	53.2	5.36	66
TCD-Kreasof	**primary**	–	–	–	–	39.9	7.93	78
	contrastive1	–	–	–	–	32.4	9.49	102
	contrastive2	–	–	–	–	21.4	14.53	335
Vicomtech	gelaco-0.25_ft_q4	20.9	9.80	38.7	8.55	41.2	7.52	37
	gelaco-0.25_ft_q8	22.0	9.27	39.0	8.42	44.4	6.75	42
	gelaco-0.50_ft_q4	18.0	12.10	31.4	10.12	31.0	9.82	94
	gelaco-0.50_ft_q8	17.9	11.57	32.2	10.15	33.7	9.24	52
	gelaco-0.75_gkd	16.1	13.90	31.8	9.93	30.6	11.03	198
	gelaco-0.75_gkd_q4	16.7	13.98	32.2	9.86	31.1	11.04	187
	gelaco-0.75_gkd_q8	15.7	13.57	31.5	9.87	31.1	10.82	197

Table 3: Translation quality metric scores on the official WMT25 GenMT test sets. XCOMET-XL and MetricX-24-Hybrid-XL scores. AyaQ and LeanAya declared support only for Czech→German. TCD-Kreasof systems did not allow to generate outputs for other languages.



(a) Total output token rates.

(b) Output rates excluding warmup times.

Figure 1: Inference speed as tokens/s when translating the entire Czech→GermanWMT25 test set. Mean and standard deviation across 3 runs, reported for various batch sizes. Primary submissions only for readability.

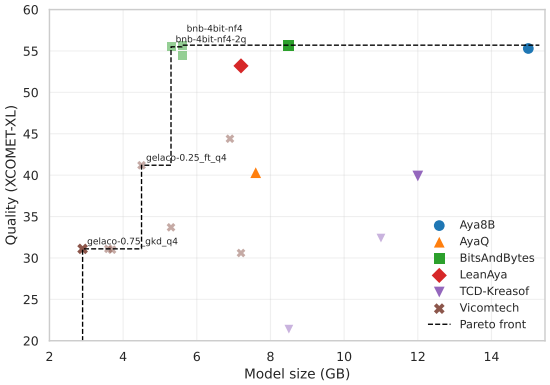
combines evolutionary layer collapse, knowledge distillation, and quantization to achieve substantial size reduction while maintaining reasonable translation performance.

Baselines As a reference system (see Table 2), we included the unmodified Aya Expanse 8B model (FP16, 16.1GB) (Dang et al., 2024) and a family of runtime–quantized variants created using the Hugging Face integration of the `bitsandbytes` library (Dettmers et al., 2022, 2023), without the use of vLLM. The baselines were not fine-tuned or adapted on the task data; their purpose was to anchor the quality–size–speed trade-offs for submitted systems. Quantized versions include 8-bit and 4-bit modes, with two 4-bit quantization data types: NF4 (normal floating point) and FP4.

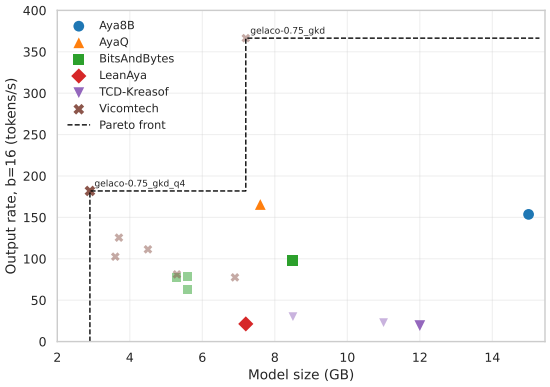
The organizers' baselines illustrate the performance one can obtain from (i) the full reference model, and (ii) straightforward, widely available post-training quantization strategies, against which more sophisticated compression pipelines can be directly compared in terms of translation quality, memory footprint, and decoding speed.

4 Results

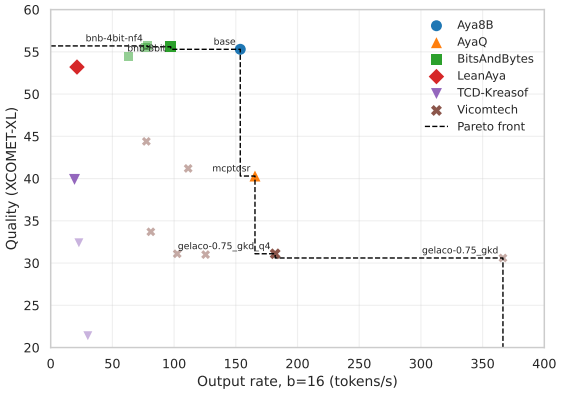
We evaluate systems' performance on the official WMT25 GenMT test sets, which comprise between 332 and 456 paragraphs for each of the three considered language pairs. Because we received submissions to the constrained track only and most



(a) Model size and COMET scores.



(b) Model size and total output token rate.



(c) Total output rate and COMET scores.

Figure 2: Comparison of translation quality, model size and inference speed for batch size 16. The staircase shows the Pareto frontline.

submissions focused on Czech→German, we primarily report results for this language pair unless mentioned otherwise.

We noticed that most of the systems may suffer from hallucinated content, which affects both translation quality and decoding speed, and complicates system comparisons. To minimize the impact of hallucinations, we segment the original paragraph-level test data at newline characters and remove empty lines. For Czech-German, this results in a test set with 2,868 segments.

To further investigate the potential impact of hallucinations, we also evaluate on a subset of the test set that only includes 1,928 segments where none of the systems exhibits hallucinations under any batch setting. This subset makes it possible to assess system performance in terms of inference speed more accurately, by excluding the distortions that hallucinations would introduce into cross-system comparisons. For brevity, analysis on the hallucination-free test data is presented in Appendix A.

Below, we briefly discuss the results in terms of quality, model size and speed. The detailed results across more benchmark settings are provided in Appendix B.

4.1 Translation Quality

Following the reference-based automatic evaluation settings proposed by the GenMT task, translation quality is automatically evaluated using XCOMET-XL[8] (Guerreiro et al., 2024) and MetricX-24-Hybrid-XL[9] (Juraska et al., 2024). Results are shown in Table 3, in which we also report the number of potentially hallucinated lines for each system, providing insight into system robustness across evaluation conditions. An output line was considered hallucinated if its output length (measured as number of characters) was at least twice the length of the source. Since we did not observe significant score differences across decoding with various batch sizes, we only report metric scores computed from outputs generated with batch size 1. However, we did observe minor differences in the outputs at different batch sizes, which indicates that the submitted implementations do not account for padding handling e.g., for relative positions (Papi et al., 2024).

The main observation is that BitsAndBytes and LeanAya achieve compression with minimal quality loss, nearly matching the baseline with ≈55 COMET scores. Other compression methods can decrease quality significantly, in particular for Czech→German, with ≈31–44 COMET scores.

4.2 Model Size

For each system variant, we record model size as the on-disk footprint of the submitted model direc-

[8] Unbabel/XCOMET-XL
[9] google/metricx-24-hybrid-xl-v2p6

Submission	System	Quality		Size↓	Speed↑ (tok/s)		
		COMET↑	MetricX↓	(GB)	b=1	b=16	b=64
Baseline	base	55.3	5.08	15.0	33.3	153.6	170.2
BitsAndBytes	**bnb-8bit**	55.6	5.08	8.5	12.3	97.3	196.4
	bnb-4bit-fp4	54.5	5.24	5.6	23.2	63.2	93.2
	bnb-4bit-nf4	55.7	5.15	5.6	23.4	78.4	134.3
	bnb-4bit-nf4-2q	55.5	5.15	5.3	19.8	76.8	131.6
AyaQ	**mcptqsr**	40.3	8.70	7.6	48.5	165.5	264.1
LeanAya	**mcsr**	53.2	5.36	7.2	2.9	21.3	27.9
TCD-Kreasof	**primary**	39.9	7.93	12.0	42.7	19.3	–
	contrastive1	32.4	9.49	11.0	50.8	22.9	–
	contrastive2	21.4	14.53	8.5	59.8	30.1	11.8
Vicomtech	gelaco-0.25_ft_q4	41.2	7.52	4.5	28.6	111.4	229.1
	gelaco-0.25_ft_q8	44.4	6.75	6.9	14.5	77.5	171.7
	gelaco-0.50_ft_q4	31.0	9.82	3.7	40.9	125.5	271.5
	gelaco-0.50_ft_q8	33.7	9.24	5.3	21.8	81.1	180.2
	gelaco-0.75_gkd	30.6	11.03	7.2	129.0	366.5	636.8
	gelaco-0.75_gkd_q4	31.1	11.04	2.9	68.3	181.9	366.5
	gelaco-0.75_gkd_q8	31.1	10.82	3.6	39.0	102.5	204.3

Table 4: Final results of the WMT25 Model Compression shared task. Primary submission names are bolded.

tory. The size is the sum of parameter shard files, tokenizer, and minimal wrapper scripts, which directly reflects the storage and transfer cost of deploying the model in gigabytes (GB).

We do not measure peak CPU memory usage or GPU VRAM footprint.

The model sizes are reported in Table 4. Organizer's submission BitsAndBytes (4-bit and 8-bit), while maintaining the translation quality, reduces the model size by up to 65%. On the other hand, Vicomtech's systems achieve best compression (81%, down to 2.9 GB from 15.0 GB) but suffer significant quality degradation.

4.3 Inference Speed

Each model was run three times per batch size, and wall-clock time was recorded for each run. Our primary metric, output rate, is defined as the number of output tokens divided by adjusted wall time (tokens/s). Output tokens were counted by retokenizing the generated hypotheses using the Aya Expanse 8B tokenizer. To isolate model initialization overhead, we performed a separate "warmup" run per model, decoding a single short sentence with batch size 1. The average wall time of warmup runs was subtracted from the total wall time to compute an adjusted speed metric. We also tested multiple batch sizes to analyze throughput scaling. The results for primary systems are presented on Figure 1.

As expected, inference speed scales with batch size, but not uniformly. Vicomtech's systems scaled most efficiently from 70 tokens/s at batch size 1 to nearly 1000 tokens/s at 512. Some models saturated early, showing minimal speedup with larger batch sizes or even failing to produce outputs.

5 Conclusion and Future Directions

The final results of the WMT25 shared task on model compression are summarized in Table 4. Figure 2 shows Pareto front comparisons across evaluation criteria.

The key findings can be summarized as follows:

- BitsAndBytes baselines and LeanAya maintained translation quality with moderate speed and model size reduction;
- Vicomtech's systems achieved best compression rates, latency and throughput thanks to efficient batch scaling, but at the cost of translation quality;
- Quantization has emerged as the most popular approach for its simplicity and effectiveness;
- Hallucinations in compressed outputs reveal the fragility of the current approaches and the need for more robust evaluation and compression-aware training techniques.

Overall, despite the moderate participation in this shared task limited the breadth of exploration, several submissions showed promising results. The results of this evaluation campaign highlight that task-specific compression of LLMs still warrants

more research efforts, especially at high compression rates required for running systems on edge devices. The smallest model still requires almost 3GB of disk space, which is incompatible with many edge devices that are equipped with a few hundred MB of memory (Cai et al., 2022). Additionally, high compression rates result in a significant performance drop. We believe that pushing the boundaries of compression rates and reducing the quality degradation in such settings represent the most interesting challenges for future research on the topic and for participants of the future editions of the task.

Looking ahead, future iterations of this task could benefit from expanding the evaluation to more language pairs, aligning more tightly with the evaluation benchmark at the GenMT task, and including human assessments of the outputs.

6 Limitations

This study offers an early glimpse into the landscape of model compression for machine translation, but several limitations constrain the generality of its findings. First, the participation was only modest, all systems compressed the same base model (Aya Expanse 8B) and primarily focused on one language pair (Czech→German). Second, the submissions relied mainly on quantization, with limited exploration of other well-established compression techniques such as parameter pruning or knowledge distillation. Third, only one system used vLLM infrastructure, limiting comparability.

Lastly, quality assessment depended solely on automatic metrics (COMET and MetricX). We did not conduct human evaluation or cross-language validation. We also did not present some important deployment metrics (e.g., memory usage, latency, energy consumption), which narrows the conclusions.

Acknowledgments

We would like to thank all participants for submitting their systems to the shared task. Marco Gaido's work has been funded by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Fortuné Kponou, Mateusz Krubiński, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Ashwin Sankar, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. 2025. Findings of the IWSLT 2025 evaluation campaign. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Red Hat AI and vLLM Project. 2024. LLM Compressor.

Alexandra Birch, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Yusuke Oda. 2018. Findings of the second workshop on neural machine translation and generation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Han Cai, Ji Lin, Yujun Lin, Zhijian Liu, Haotian Tang, Hanrui Wang, Ligeng Zhu, and Song Han. 2022. Enable deep learning on mobile devices: Methods, systems, and applications. *ACM Trans. Des. Autom. Electron. Syst.*, 27(3).

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Jacob Devlin. 2017. Sharp models on dull hardware: Fast and accurate neural machine translation decoding on the CPU. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2820–2825, Copenhagen, Denmark. Association for Computational Linguistics.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. 2020. Stabilizing the lottery ticket hypothesis.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. In *The Eleventh International Conference on Learning Representations*.

Thamme Gowda, Roman Grundkiewicz, Elijah Rippeth, Matt Post, and Marcin Junczys-Dowmunt. 2024. PyMarian: Fast neural machine translation and evaluation in python. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 328–335, Miami, Florida, USA. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. Findings of the third workshop on neural generation and translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, Hong Kong. Association for Computational Linguistics.

Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch. 2020. Findings of the fourth workshop on neural generation and translation. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 1–9, Online. Association for Computational Linguistics.

Kenneth Heafield, Biao Zhang, Graeme Nail, Jelmer Van Der Linde, and Nikolay Bogoychev. 2022. Findings of the WMT 2022 shared task on efficient translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 100–108, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kenneth Heafield, Qianqian Zhu, and Roman Grundkiewicz. 2021. Findings of the WMT 2021 shared task on efficient translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 639–651, Online. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Preliminary ranking of wmt25 general machine translation systems.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yasmin Moslem, Muhammad Hazim Al Farouq, and D. John Kelleher. 2025. Iterative Layer Pruning for Efficient Translation Inference. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China.

Sara Papi, Marco Gaido, Andrea Pilzer, and Matteo Negri. 2024. When good and reproducible results are a giant with feet of clay: The importance of software quality in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3657–3672, Bangkok, Thailand. Association for Computational Linguistics.

David Ponce, Thierry Etchegoyhen, and Javier Del Ser. 2025a. Gelaco: An evolutionary approach to layer compression. *arXiv preprint arXiv:2507.10059*.

David Ponce, Harritxu Gete, and Thierry Etchegoyhen. 2025b. Vicomtech@WMT 2025: Evolutionary Model Compression for Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang, and Jie Tang. 2023. GKD: A general knowledge distillation framework for large-scale pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 134–148, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang and Anshumali Shrivastava. 2025. LeanQuant: Accurate and Scalable Large Language Model Quantization with Loss-error-aware Grid. In *International Conference on Representation Learning*, volume 2025, pages 35521–35544.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

A Hallucinations

To understand if the potential hallucinations impacted the results, we benchmarked the participating systems on a subset of the Czech→German WMT25 test set. This subset includes only segments where none of the systems exhibits hallucinations under any batch setting. An output line was considered hallucinated if its tokenized output length was at least twice the length of the tokenized source. This version of the Czech-German test set reduces the number of input segments from 2,686 to 1,928 segments. Table 5 illustrates quality comparisons across systems for both versions of the test set using a reference-less metric, WMT23-CometKiwi-XL (Rei et al., 2023), computed using Pymarian (Gowda et al., 2024). Inference speed metrics across different settings are presented in Figures 3 and 4.

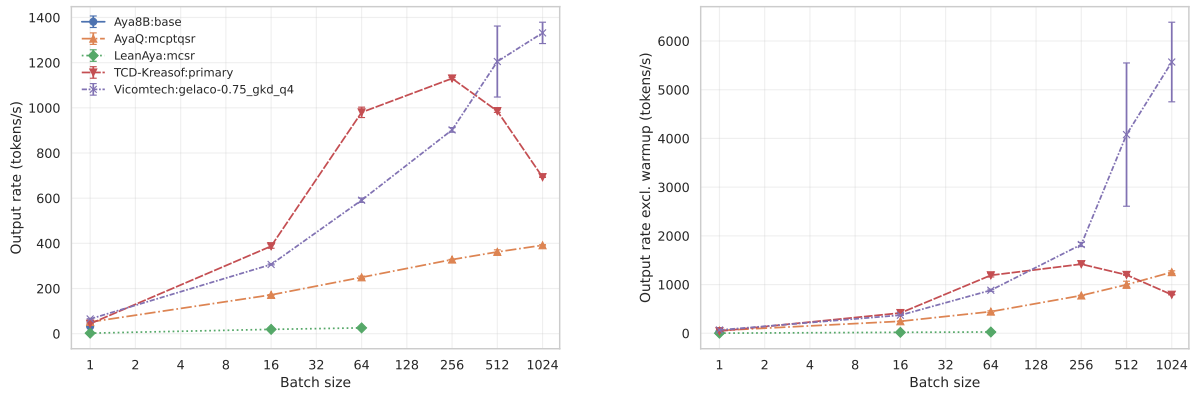
Submission	**System**	*full* **CKiwi↑**	*subset* **CKiwi↑**	**#Halluc.**
Baseline	base	66.9	71.1	83
BitsAndBytes	**bnb-8bit**	66.8	—	86
	bnb-4bit-fp4	66.5	70.8	153
	bnb-4bit-nf4	66.6	70.7	103
	bnb-4bit-nf4-2q	66.7	70.8	106
AyaQ	**mcptqsr**	58.9	64.9	200
LeanAya	**mcsr**	66.3	70.9	66
TCD-Kreasof	**primary**	59.7	64.4	78
	contrastive1	55.0	58.8	102
	contrastive2	39.3	42.9	335
Vicomtech	gelaco-0.25_ft_q4	60.5	64.1	37
	gelaco-0.25_ft_q8	61.8	65.4	42
	gelaco-0.50_ft_q4	53.3	56.7	94
	gelaco-0.50_ft_q8	55.9	59.3	52
	gelaco-0.75_gkd	54.8	59.3	198
	gelaco-0.75_gkd_q4	54.5	59.9	187
	gelaco-0.75_gkd_q8	55.2	60.6	197

Table 5: COMET-Kiwi-XL scores for the original Czech→German WMT25 test set and the filtered version without lines exhibiting potential hallucinations.

B Detailed results

Figure 5 presents extended evaluation of the average output token rates across multiple batch sizes for all submissions, including contrastive submissions.

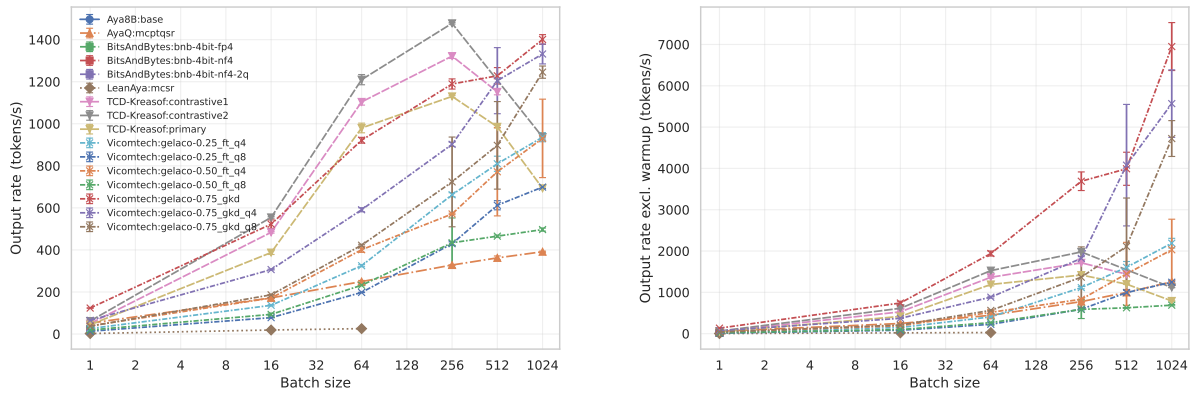
Table 6 provides details about warmup times and total decoding times for two batch size settings for each system.



(a) Total output rates.

(b) Output rates excluding warmup times.

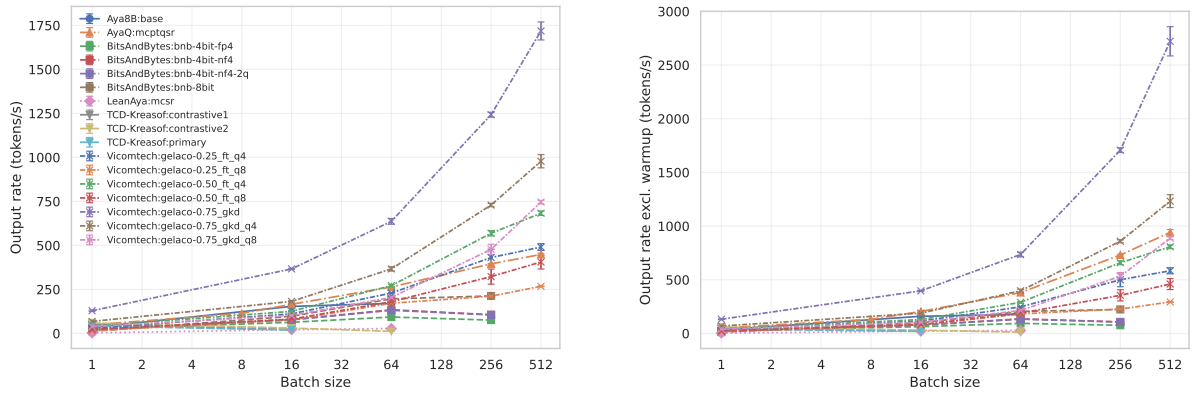
Figure 3: Inference speed as tokens/s when translating **the subset of the Czech→German WMT25 test set not causing hallucinations**. Mean and standard deviation across 3 runs, reported for various batch sizes. Primary submissions only.



(a) Total output rates.

(b) Output rates excluding warmup times.

Figure 4: Inference speed as tokens/s when translating **the subset of the Czech→German WMT25 test set not causing hallucinations**. Mean and standard deviation across 3 runs, reported for various batch sizes. Primary and contrastive submissions.



(a) Total output rates.

(b) Output rates excluding warmup times.

Figure 5: Inference speed as tokens/s when translating the entire Czech→German WMT25 test set. Mean and standard deviation across 3 runs, reported for various batch sizes. **Primary and contrastive submissions**.

Submission	System	Warmup (sec.)	Time (sec.)	Speed↑ (tok/s) total	excl.w.	Time (sec.)	Speed↑ (tok/s) total	excl.w.
			Batch size 1			Batch size 16		
Baseline	base	22.5	2,659.3	33.3	33.6	576.6	153.6	159.9
BitsAndBytes	**bnb-8bit**	19.7	7,231.4	12.3	12.3	923.2	97.3	99.4
	bnb-4bit-fp4	7.6	4,065.9	23.2	23.2	1,492.0	63.2	63.5
	bnb-4bit-nf4	7.4	3,762.6	23.4	23.5	1,123.1	78.4	78.9
	bnb-4bit-nf4-2q	7.8	4,461.4	19.8	19.8	1,148.7	76.8	77.3
AyaQ	**mcptqsr**	62.5	1,111.5	48.5	51.4	325.9	165.5	204.8
LeanAya	**mcsr**	25.7	32,198.5	2.9	2.9	4,291.8	21.3	21.4
TCD-Kreasof	**primary**	10.1	4,228.0	42.7	42.8	8,931.0	19.3	19.3
	contrastive1	9.2	4,033.2	50.8	50.9	8,970.6	22.9	22.9
	contrastive2	8.5	2,994.1	59.8	60.0	5,412.6	30.1	30.2
Vicomtech	gelaco-0.25_ft_q4	25.6	2,711.9	28.6	28.9	729.6	111.4	115.4
	gelaco-0.25_ft_q8	26.4	5,296.0	14.5	14.6	1,015.7	77.5	79.6
	gelaco-0.50_ft_q4	25.6	2,683.6	40.9	41.3	833.1	125.5	129.5
	gelaco-0.50_ft_q8	25.3	4,083.0	21.8	22.0	1,155.5	81.1	83.0
	gelaco-0.75_gkd	26.0	942.1	129.0	132.7	339.2	366.5	396.9
	gelaco-0.75_gkd_q4	25.4	1,773.0	68.3	69.3	670.6	181.9	189.1
	gelaco-0.75_gkd_q8	26.4	3,201.9	39.0	39.3	1,201.4	102.5	104.8

(a) Speed metrics for **batch sizes 1 and 16**.

Submission	System	Warmup (sec.)	Time (sec.)	Speed↑ (tok/s) total	excl.w.	Time (sec.)	Speed↑ (tok/s) total	excl.w.
			Batch size 64			Batch size 256		
Baseline	base	22.5	520.5	170.2	177.9	–	–	–
BitsAndBytes	**bnb-8bit**	19.7	456.6	196.4	205.3	419.1	213.0	223.5
	bnb-4bit-fp4	7.6	1,011.3	93.2	93.9	1,245.5	75.6	76.1
	bnb-4bit-nf4	7.4	655.5	134.3	135.8	825.8	106.6	107.6
	bnb-4bit-nf4-2q	7.8	670.3	131.6	133.1	839.5	105.0	106.0
AyaQ	**mcptqsr**	62.5	204.0	264.1	380.8	136.0	395.2	731.0
LeanAya	**mcsr**	25.7	3,293.0	27.9	28.1	–	–	–
TCD-Kreasof	**primary**	10.1	–	–	–	–	–	–
	contrastive1	9.2	–	–	–	–	–	–
	contrastive2	8.5	12,428.5	11.8	11.8	–	–	–
Vicomtech	gelaco-0.25_ft_q4	25.6	346.0	229.1	247.4	188.6	431.1	500.2
	gelaco-0.25_ft_q8	26.4	455.9	171.7	182.2	384.7	211.5	227.1
	gelaco-0.50_ft_q4	25.6	403.0	271.5	290.0	188.0	569.0	658.7
	gelaco-0.50_ft_q8	25.3	518.0	180.2	189.5	288.5	322.8	354.9
	gelaco-0.75_gkd	26.0	193.8	636.8	735.7	95.6	1,242.6	1,707.5
	gelaco-0.75_gkd_q4	25.4	327.3	366.5	397.4	168.5	729.3	858.9
	gelaco-0.75_gkd_q8	26.4	582.1	204.3	214.0	255.6	477.5	532.6

(b) Speed metrics for **batch sizes 64 and 256**.

Table 6: Detailed speed metrics including the total translation time of the Czech→German WMT25 test set, the total token output rate and token output rate excluding warmup. Averages across 3 runs for various batch sizes.

Findings of the WMT 2025 Shared Task of the Open Language Data Initiative

David Dale* Meta FAIR
Laurie Burchell* Common Crawl Foundation
Jean Maillard Meta FAIR

Idris Abdulmumin University of Pretoria
Antonios Anastasopoulos George Mason University

Isaac Caswell Google
Philipp Koehn Johns Hopkins University

Correspondence: info@oldi.org

Abstract

We present the results of the WMT 2025 shared task of the Open Language Data Initiative (OLDI). Participants were invited to contribute to the existing massively multilingual open datasets supported by OLDI (FLORES+, OLDI-Seed, WMT24++), or create new resources in line with OLDI's aims. We accepted eight submissions: seven extensions or revisions of the existing datasets and one submission with a new massively parallel training dataset, SMOL. These contributions advance the coverage and quality of multilingual datasets, especially since for many languages, they are the first publicly-available training or evaluation data for machine translation. All contributions are released under permissive open-source licenses.

1 Introduction

Recent advances in machine translation (MT) have resulted in system output which is indistinguishable from that of human translators (Kocmi et al., 2024). However, even if we assume that the task of MT is 'solved', this would be true only for a small number of well-resourced language pairs. Achieving such high task performance requires abundant parallel training data specific to the language pair and domain of interest, either as explicit parallel datasets, or as incidental bilingual signals in generic web data (Briakou et al., 2023). For the majority of the world's languages, we lack both the parallel training data necessary to train MT models, as well as the evaluation data to assess their translation capabilities.

One way to address this bottleneck is creation of massively multilingual parallel datasets and their extension to new languages. In this paper, we describe the second Shared Task for Open Language Data Initiative (OLDI) which invited language communities to contribute to high-quality, massively parallel and open-source datasets. Contributions could involve either extending existing datasets to new language varieties, making substantial improvements to existing datasets, or creating new massively multilingual parallel datasets. The datasets of interest to the shared task include (but are not limited to) FLORES+, OLDI-Seed (NLLB Team et al., 2024; Maillard et al., 2024), and WMT24++ (Deutsch et al., 2025). The OLDI itself is a community of researchers that maintains the former two datasets and promotes better language resources for under-served languages in general.[1]

This year, we received eight submissions, including six extensions of datasets to new language varieties, two revisions of existing translations, and one entirely new massively parallel dataset. All the data will be made available online under permissive open-source licenses.[2]

2 Datasets

2.1 FLORES+

FLORES is a family of datasets designed to benchmark multilingual translation, with many-to-many alignment across over 200 languages. The first iteration of this dataset covered only three languages (Guzmán et al., 2019), but following iterations increased coverage first to 101 languages (FLORES-101, Goyal et al., 2022) and then to over 200 languages as part of the "No Language Left Behind" project (NLLB Team et al., 2024). Finally, as part

*Equal contribution

[1] https://oldi.org/

[2] https://huggingface.co/collections/openlanguagedata

Proceedings of the Tenth Conference on Machine Translation, pages 495–502

Language	Variety	ISO 639-3	ISO 15924	Glottocode	Contributors	Contributions
123 languages					Caswell et al. (2025)	SMOL (new dataset)
Ladin	Val Badia	lld	Latn	badi1244	Frontull et al. (2025)	FLORES+ new data
	Gherdëina			gard1241		
Kyrgyz		kir	Cyrl	kirg1245	Jumashev et al. (2025)	OLDI-Seed new data
Norvegian Bokmål	moderate	nob	Latn	under review	Mæhlum et al. (2025)	FLORES+ revision
	radical			under review		FLORES+ new data
Southern Uzbek		uzs	Arab	sout2699	Mamasaidov et al. (2025)	FLORES+ new data (dev split)
French		fra	Latn	stan1290	Marmonier et al. (2025)	OLDI-Seed new data
Standard Moroccan Tamazight		zgh	Tfng	stan1324	Oktem et al. (2025)	FLORES+ revision, OLDI-Seed revision
Romansh	Rumantsch Grischun	roh	Latn	ruma1247	Vamvas et al. (2025)	WMT24++ new data
	Sursilvan			surs1244		
	Surmiran			surm1243		
	Sutsilvan			suts1235		
	Puter			uppe1396		
	Vallader			lowe1386		

Table 1: A summary of all contributions to the WMT 2025 Shared Task of the Open Language Data Initiative.

of the previous edition of this shared task, an additional 8 languages were included on top of several corrections to existing datasets (Abdulmumin et al., 2024; Ali et al., 2024; Gordeev et al., 2024; Kuzhuget et al., 2024; Mamasaidov and Shopulatov, 2024; Perez-Ortiz et al., 2024; Yu et al., 2024). This new, living version of the FLORES benchmark is released under the name FLORES+.

2.2 OLDI-Seed

The NLLB-Seed dataset of NLLB Team et al. (2024) was created as a source of starter data for languages without publicly-available high-quality bitext in sufficient quantity for training natural language processing (NLP) models. This dataset consists of around 6000 sentences sampled from the Wikipedia articles listed in English Wikimedia's "List of articles every Wikipedia should have".[3] These were professionally translated into each of the 38 languages covered by the first iteration of this dataset (39 if including English), and experiments by Maillard et al. (2023) demonstrated the gains of including these datasets in the training mix of MT models.

[3] https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have

Participants to last year's edition of this shared task contributed three new languages (Ahmed et al., 2024; Cols, 2024; Ferrante, 2024). To reflect the continuously updating nature of this dataset, and to distinguish it from prior iterations, it is released as OLDI-Seed.

2.3 WMT24++

The WMT24++ dataset (Deutsch et al., 2025) was created by translating the test dataset from the WMT24 General MT shared task (Kocmi et al., 2024) from English to 54 other languages. The 998 paragraph-sized English source documents come from four different domains: literary, news, social, and transcribed speech. Thus, WMT24++ is mostly complementary to FLORES+ in document sizes and domains (though news is an overlapping domain). Unlike the two previous datasets, WMT24++ is managed by Google Research and not by OLDI, but in common with all datasets promoted by OLDI, it is released under a permissive license (Apache License 2.0).

2.4 Other datasets

There are other massively parallel datasets that could have been potential targets for extension

in the OLDI shared task. They include MT evaluation benchmarks such as NTREX-128 (Federmann et al., 2022) and BOUQuET (Andrews et al., 2025), as well as other parallel datasets that could be reused for MT like Global MMLU (Singh et al., 2025) or MCS-350 (Agarwal et al., 2023). FLEURS, a parallel datasets of speech (Conneau et al., 2023) and signed language (Tanzer, 2025; Costa-jussà et al., 2025) could also have been considered. Finally, there is the massively parallel GATITOS dataset (Jones et al., 2023) of 4000 frequently used words and phrases translated from English that served as a foundation for SMOL (Caswell et al., 2025), one of the contributions of the current shared task.

3 Shared task definition

The goal of the shared task was to expand high-quality, massively-parallel and open-source datasets to improve the resources available for multilingual applications like MT. Contributions could consist of the addition of new language varieties to existing datasets, substantial improvements to existing datasets, or novel datasets compatible with the aims of OLDI. Most contributions were to the datasets managed by OLDI: FLORES+ and OLDI-Seed.

3.1 Contributing to FLORES+ and OLDI-Seed

We encouraged the contributors of new languages to FLORES+ and OLDI-Seed to start from the original English data; using a different pivot language was also possible, if clearly documented. We required the translations to be performed, wherever possible, by qualified, native speakers of the target language, and encouraged verification of the data by at least one additional native speaker. More recommendations were described in the OLDI contribution guidelines.[4]

For FLORES+ translation, we did not allow using or even referencing MT output, including post-editing, to avoid introducing any machine bias in this evaluation dataset. For OLDI-Seed data, the use of post-edited machine translated content was allowed, as long as all data was manually verified and the MT system allowed reusing their outputs to train other models (which is not the case for the major commercial LLMs). This is because OLDI-Seed is intended as MT training data rather than evaluation data and so is subject to less strict requirements.

We asked the participants to attach dataset cards to new data submissions, detailing precise language information and the translation workflow that was employed. In particular, we asked them to identify the language with both an ISO 639-3 individual language tag and a Glottocode, and identify the script with an ISO 15924 script code. For example, the Rumantsch Grischun variety was identified as `roh_Latn_ruma1247`.

Participants were encouraged to provide experimental validation of the quality of the data they were submitting.

3.2 Contributing other data

We also accepted extensions and improvements to other foundational multilingual datasets (e.g. WMT24+) that are massively parallel, open source, and useful to under-served language communities. We suggested that contribution workflow should follow that for FLORES and OLDI-Seed as closely as possible to ensure data quality and documentation. We required contributed data to be released under an open license (allowing free research use as a minimum).

4 Submissions

4.1 Shared task submissions

Table 1 summarises the contributions accepted as part of the shared task and the languages that were involved. In the rest of this section, we briefly describe each submission.

Caswell et al. (2025) created the SMOL dataset: a multiway parallel training dataset with high lexical coverage. The first part of the dataset, SMOLSENT, is based on 863 English sentences semi-manually selected from Common Crawl data[5] to cover 5.5k of the most common English words (obtained by joining the GATITOS wordlist and the most frequent words in Common Crawl). The second part of the dataset, SMOLDOC, is based on 584 English documents generated with LLMs using prompt templates that ensured diversity of topics and styles. The dataset was professionally translated from English into 115 languages, mostly under-resourced. Subsequently, additional volunteer translations were contributed, bringing the total number of languages to 123.

[4] https://oldi.org/guidelines

[5] https://commoncrawl.org/

To demonstrate the value of the dataset, the authors used it for in-context learning of several commercial LLMs and for fine-tuning of a GEMINI LLM for translation out of English into the 80 languages for which evaluation data were available. For most language subsets and models, in-context learning with SMOL examples was found to be superior to zero-shot translation. Fine-tuning demonstrated positive effect of both SMOL dataset parts and their combination with GATITOS.

Frontull et al. (2025) translated FLORES+ into Val Badia and Gherdëina, two varieties of the Ladin language which is spoken in Northern Italy. The paper gives a detailed overview of Ladin and the resources available for it. The FLORES sentences were first manually translated into the Val Badia variety, using German, Italian, Friulian, and English references, then into the Gherdëina variant, using Val Badia as an additional reference. The authors additionally released training datasets for Gherdëina–Italian and Val Badia–Gherdëina pairs and used them to fine-tune an NLLB model to translate between the three languages. They used the newly translated FLORES dataset to benchmark the MT performance of this model and four LLMs (with and without retrieval of few-shot examples from the parallel training dataset). They found that even though retrieval helps, translation into the Ladin variants remains a clear challenge for current LLMs.

Jumashev et al. (2025) expand OLDI-Seed to Kyrgyz by post-editing LLM-based translations from English (using also Kazakh and Russian lexical resources) with a subsequent review to ensure term consistency throughout the dataset. Two post-editing techniques that the authors highlight are breaking a complex English sentence into two or more Kyrgyz sentences to be more fluent under the Kyrgyz SOV sentence structure, and a careful choice between native Kyrgyz words and Russian or English calques for scientific terms.

To demonstrate the effectiveness of the resulting parallel dataset, the authors finetuned four multilingual models on it and demonstrate gains in translation performance of each model on FLORES+ and X-WMT (Mirzakhalov et al., 2021).

Mæhlum et al. (2025) revise the FLORES+ dataset in Norvegian Bokmål and create a new version of it in Radical Bokmål, a sub-variety that is closer to spoken Norwegian dialects than the more Danish-like conservative Bokmål that dominates formal discourse. The authors provide a detailed explanation of the difference between the varieties, followed by an overview of the grammatical and lexical mistakes present in the original Bokmål FLORES+ dataset, such as anglicisms, word-by-word translations and problems in agreement. The necessary revisions affected two thirds of the FLORES+ sentences. The authors demonstrate that the new version of the dataset, cleaned from anglicisms and overly literal translations, serves as a more challenging reference set for English-Bokmål translation than the previous version.

Mamasaidov et al. (2025) extended FLORES+ to Southern Uzbek, a variety spoken in Afghanistan and written in Arabic script. It is substantially different from Northern Uzbek, which is spoken in Uzbekistan and written in Latin. The challenges of understanding and generating Southern Uzbek include the ambiguity of Arabic vowel characters and the use of a zero-width non-joiner character (U+200C) to separate the words' suffixes. Apart from the FLORES+ dev set translation into Southern Uzbek performed by a single native linguist, the paper also contributes an automatically aligned parallel dataset of the Southern and Northern Uzbek sentences, a NLLB model fine-tuned with this data and evaluated with FLORES+, and scripts for transliteration of Southern Uzbek into Latin and for post-correction of missing U+200C characters. The newly finetuned model outperforms the strong LLM baselines on translation into Southern Uzbek, demonstrating the lack of previous support for this language.

Marmonier et al. (2025) expand OLDI-Seed to French with the purpose of serving as a pivot language for the under-resourced regional languages of France. Each OLDI-Seed sentence has been translated from English with 9 different MT systems, and two native French speakers selected and post-edited the most promising translation candidate from each such set. Finally, the translations were processed through a grammar checker. For validating the post-edited translations, the authors use MetricX-24 quality estimation system (Juraska et al., 2024), demonstrating that the human translations result in lower predicted error rates than any of the MT candidates. The paper emphasizes the terminological complexity of the OLDI-Seed dataset and the challenges of producing fluent French translations as a result of the issues sometimes found in the English source sentences.

Oktem et al. (2025) revised FLORES+ and OLDI-Seed sentences in Standard Moroccan

Tamazight as a part of the Awal initiative. The FLORES sentences were revised by two linguists using English as reference whilst OLDI-Seed was revised by three professional Tamazight translators with English and Arabic references. 36% of FLORES and overall and 40% of OLDI-Seed sentences required correction of spelling mistakes, transliteration errors, unnecessary or malformed loanwords, and mistranslations.

The authors fine-tuned an NLLB-based model with the corrected OLDI-Seed dataset and other Tamazight-English parallel datasets and evaluated it alongside with the original NLLB models and commercial LLMs on the original and corrected FLORES dataset. They found that the corrected FLORES dataset yields better MT evaluation metrics, and that fine-tuning with the OLDI-Seed data improves NLLB performance, making the model outperform the LLMs in the English-Tamazight direction.

Vamvas et al. (2025) expanded the WMT24++ benchmark with six varieties of the Romansh language: Rumantsch Grischun, a supra-regional variety, and five regional varieties: Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. The benchmark texts were translated from German by hired professionals who are native speakers of both German and a Romansh variety. The translations were then reviewed by two expert linguists. For automatic validation of the translations, the authors used language identification with a FastText model and cross-variety ChrF++ scores, demonstrating that the texts in the Romansh varieties are similar but distinguishable from each other. The resulting benchmark was used to assess the performance of MT system and LLMs on translation between German and Romansh, demonstrating that although some models already understand Romansh fairly well, translation into it is still challenging.

4.2 Other dataset extensions

It should be noted that not all contributors to the OLDI datasets submitted shared task papers. In the last year, FLORES+ has also received new translations in Chuvash, Dargwa, and Meadow Mari, regional languages in Russia, and incorporated the translations into Nko (Doumbouya et al., 2023) and five Indic languages (Gala et al., 2023): Bodo, Dogri, Konkani, Sindhi and Manipuri. OLDI-Seed has been extended with the Nko language (Doumbouya et al., 2023).[6]

5 Discussion

Creating and maintaining language resources and technologies is hard, especially massively multilingual ones. There are tradeoffs and compromises: between the number of language varieties covered and the depth of the support of each variety, between the difficulty of benchmarks and the ease of translating them into new languages, between the naturalness of the translation in the target language and its faithfulness to the source content. Without the active interest of the communities actually speaking the language, advancing the NLP technologies for many of the world's under-served languages is hardly possible.

The contributions of this year's OLDI shared task highlight some of the issues with existing multilingual datasets and put forward suggestions as to how these might be solved. One issue which was highlighted repeatedly by the teams translating OLDI-Seed is its terminological complexity and the requirement of specialized knowledge for translating it. However, the emergence of SMOL as an alternative seed training dataset for MT helps circumvent this issue. An additional issue is that many popular multilingual datasets are English-centric. This is a barrier for extension into languages whose speakers use other languages as a lingua franca. Contributions like French OLDI-Seed by Marmonier et al. (2025) mitigate this. Finally, the work of Mæhlum et al. (2025) and Oktem et al. (2025) on revising the OLDI datasets in Norvegian Bokmål and Tamazight, respectively, show the need for continuous improvement of massive parallel datasets, especially with the direct involvement of the community of speakers.

Since the previous round of the OLDI shared task, contributions to FLORES+ and OLDI-Seed have already propagated to massively multilingual NLP benchmarks (e.g. Luo et al. (2025)) and to the extension of foundation models to new languages (e.g. Tsiamas et al. (2025)). We hope that the new datasets, languages and revisions contributed in the current shared task will similarly lead to further improvements in NLP resources and MT research for under-resourced languages.

[6] See the detailed list of changes and their attributions in the `CHANGELOG.md` files and dataset cards in the FLORES+ and OLDI-Seed repositories.

6 Conclusion

We presented the results of the WMT 2025 OLDI shared task. We accepted 8 submissions covering 16 languages, including the new SMOL dataset covering 123 languages, and extensions or revisions of the existing foundational datasets, FLORES+, OLDI-Seed, and WMT24++, in 14 language varieties. We are truly grateful to all participants for their work and we hope that these contributions are soon adopted by the research community, enhancing a positive feedback loop between the developers of language technologies and the communities of language speakers.

Acknowledgments

OLDI functions on volunteer time and community contributions. We are grateful to the language communities, researchers, and reviewers that contributed to this shared task with new resources for under-served languages. AA also acknowledges support from the US National Science Foundation under award CIRC-2346334.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Firoz Ahmed, Nitin Venkateswaran, and Sarah Moeller. 2024. The Bangla/Bengali seed dataset submission to the WMT24 open language data initiative shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 556–566, Miami, Florida, USA. Association for Computational Linguistics.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. Association for Computational Linguistics.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, and 1 others. 2025. BOUQuET: dataset, benchmark and open initiative for universal quality evaluation in translation. *arXiv preprint arXiv:2502.04314*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Moussa Koulako Bala Doumbouya, Djibrila Diane, Baba Mamadi Diane, Solo Farabado, Edoardo Ferrante, Alessandro Guasoni, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, and 3 others. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 740–760, Suzhou, China. Association for Computational Linguistics.

Jose Cols. 2024. Spanish corpus and provenance with computer-aided translation for the WMT24 OLDI shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 624–635, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Marta R. Costa-jussà, Bokai Yu, Pierre Andrews, Belen Alastruey, Necati Cihan Camgoz, Joe Chuang, Jean Maillard, Christophe Ropers, Arina Turkatenko, and Carleigh Wood. 2025. 2M-BELEBELE: Highly multilingual speech and American Sign Language comprehension dataset download PDF. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10893–10904, Vienna, Austria. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *Preprint*, arXiv:2502.12404.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye

Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory Conde, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine translation for nko: Tools, corpora, and baseline results. In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Edoardo Ferrante. 2024. A high-quality seed dataset for Italian machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 567–569, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto, and Paolo Anvidalfarei. 2025. Bringing ladin to flores+. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 698–708, Suzhou, China. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Isai Gordeev, Sergey Kuldin, and David Dale. 2024. FLORES+ translation and machine translation evaluation for the Erzya language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Murat Jumashev, Alina Tillabaeva, Aida Kasieva, Turgunbek Omurkanov, Akylai Musaeva, Meerim Emil kyzy, Gulaiym Chagataeva, and Jonathan North Washington. 2025. The kyrgyz seed dataset submission to the wmt25 open language data initiative shared task. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 725–739, Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Ali Kuzhuget, Airana Mongush, and Nachyn-Enkhedorzhu Oorzhak. 2024. Enhancing tuvan language resources through the FLORES dataset. In *Proceedings of the Ninth Conference on Machine Translation*, pages 593–599, Miami, Florida, USA. Association for Computational Linguistics.

Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, and 1 others. 2025. Gloteval: A test suite for massively multilingual evaluation of large language models. *arXiv preprint arXiv:2504.04155*.

Jean Maillard, Laurie Burchell, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. Findings of the WMT 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Mukhammadsaid Mamasaidov, Azizullah Aral, Abror Shopulatov, and Mironshoh Inomjonov. 2025. Filling the gap for uzbek: Creating translation resources for southern uzbek. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 718–724, Suzhou, China. Association for Computational Linguistics.

Mukhammadsaid Mamasaidov and Abror Shopulatov. 2024. Open language data initiative: Advancing low-resource machine translation for Karakalpak. In *Proceedings of the Ninth Conference on Machine Translation*, pages 606–613, Miami, Florida, USA. Association for Computational Linguistics.

Malik Marmonier, Benoît Sagot, and Rachel Bawden. 2025. A french version of the oldi seed corpus. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 685–697, Suzhou, China. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Bekhzodbek Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, Francis Tyers, Orhan Firat, John Licato, and Sriram Chellappan. 2021. Evaluating multiway multilingual NMT in the Turkic languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 518–530, Online. Association for Computational Linguistics.

Petter Mæhlum, Anders Næss Evensen, and Yves Scherrer. 2025. Improved norwegian bokmål translations for flores. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 761–769, Suzhou, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Alp Oktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja, and Farida Boudichat. 2025. Correcting the tamazight portions of flores+ and oldi seed datasets. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 709–717, Suzhou, China. Association for Computational Linguistics.

Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusèp Loís Sans Socasau, and Juan Pablo Martínez. 2024. Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian. In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Garrett Tanzer. 2025. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6167–6191, Albuquerque, New Mexico. Association for Computational Linguistics.

Ioannis Tsiamas, David Dale, and Marta R. Costa-jussà. 2025. Improving language and modality transfer in translation by character-level modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20171–20187, Vienna, Austria. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Battesta Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna Rutkiewicz, and Rico Sennrich. 2025. Expanding the wmt24++ benchmark with rumantsch grischun, sursilvan, sutsilvan, surmiran, puter, and vallader. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 665–684, Suzhou, China. Association for Computational Linguistics.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu Chinese: Workflow and analysis. In *Proceedings of the Ninth Conference on Machine Translation*, pages 600–605, Miami, Florida, USA. Association for Computational Linguistics.

Findings of the WMT 2025 Shared Task LLMs with Limited Resources for Slavic Languages: MT and QA

Shu Okabe[1,2] **Daryna Dementieva**[1,2] **Marion Di Marco**[1,2] **Lukas Edman**[1,2]
Kathy Hämmerl[1,2] **Marko Měškank**[3] **Anita Hendrichowa**[3] **Alexander Fraser**[1,2,4]

[1]Technische Universität München [2]Munich Center for Machine Learning
[3]WITAJ-Sprachzentrum [4]Munich Data Science Institute

Correspondence: shu.okabe@tum.de, daryna.dementieva@tum.de

Abstract

We present the findings of the WMT 2025 Shared Task *LLMs with Limited Resources for Slavic Languages*. This shared task focuses on training LLMs using limited data and compute resources for three Slavic languages: Upper Sorbian (hsb), Lower Sorbian (dsb), and Ukrainian (uk), with the objective to develop and improve LLMs for these languages. We consider two tasks which are to be evaluated jointly: Machine Translation (MT) and Multiple-Choice Question Answering (QA).

In total, three teams participated in this shared task, with submissions from all three teams for the Sorbian languages and one submission for Ukrainian. All submissions led to an improvement compared to the baseline Qwen2.5-3B model through varying fine-tuning strategies. We note, however, that training purely on MT degrades original QA capabilities. We also report further analyses on the submissions, including MT evaluation using advanced neural metrics for Ukrainian, as well as manual annotation and comparison to the current Sorbian machine translator.

1 Introduction

For a large majority of the world's languages, only limited resources are available for training NLP tools, but modern large language models (LLMs) need large amounts of both labelled and unlabelled data to function well. Improving the coverage of low-resource languages in LLMs is an active research area. Recent examples include Nag et al. (2025) for Indic languages and Tonja et al. (2024) for Ethiopian languages. Similarly, there is active work on low-resource machine translation, with recent shared tasks and datasets covering translation for low-resource Indic languages (Pakray et al., 2024), Creole languages (Robinson et al., 2024) as well as Indigenous Languages of the Americas (De Gibert et al., 2025).

Although commercial LLMs increasingly show high performance on both general tasks and machine translation, specialised MT models are still typically required for best results. Our challenge to participants in this shared task is to build a model under low-resource conditions to jointly optimise machine translation and question answering. We aim to study potential synergy effects between these two tasks, as well as to explore whether optimising for one task in a low-resource setting will negatively impact the other task. We are one of the first WMT shared tasks to focus on joint optimisation of Machine Translation (MT) and Question Answering (QA). The Multilingual Instruction Shared Task in the same year took a similar approach, but allowed significantly larger models.

Our task focuses on three Slavic languages: Upper Sorbian, Lower Sorbian, and Ukrainian. Thus, we aim to highlight both truly low-resource settings and mid-resource language scenarios in the context of modern language technologies. As our goal is to evaluate LLMs as general-purpose tools for a given language, we designed our setup to mirror widely adopted benchmarks such as GLUE (Wang et al., 2019) and MMLU (Hendrycks et al., 2021). Since Sorbian languages and Ukrainian currently lack such comprehensive language understanding benchmarks, we approximated a multitask evaluation by selecting two representative tasks: Machine Translation and Question Answering.

Previous iterations of the WMT Shared Tasks on translating low-resource languages (Weller-Di Marco and Fraser, 2022; Libovický and Fraser, 2021; Fraser, 2020) compared supervised and unsupervised translation in various data settings. For both Sorbian languages, the WITAJ-Sprachzentrum provided new machine translation and question answering datasets for this shared task. For Ukrainian, the MT portion of the dataset corresponds to that of the WMT 2025 general translation

Proceedings of the Tenth Conference on Machine Translation, pages 503–519
November 8-9, 2025

task,[1] and the QA portion is based on the dataset from the UNLP 2024 Shared Task on fine-tuning LLMs for Ukrainian (Romanyshyn et al., 2024).

Our research questions are as follows:

- In a low-resource scenario, how does training a model for machine translation impact its performance on a secondary task such as question answering?
- Is it possible to improve capabilities on both machine translation and question-answering in a small LLM?

This article is structured as follows: Section 2 describes the shared task rules, while Section 3 details the MT and QA datasets provided for both development and test phases. Section 4 presents the systems devised by the three participating teams. Section 5 displays the official leaderboard for the primary submissions in all three tracks. Section 6 analyses the model outputs with some additional experiments.

2 Shared Task Description

2.1 Languages

Upper Sorbian (ISO code: `hsb`; Glottocode:[2] `uppe1395`) and Lower Sorbian (`dsb`; `lowe1385`) are minority languages spoken in the eastern part of Germany in the federal states of Saxony and Brandenburg, with only 30k and 7k native speakers, respectively. As western Slavic languages, Upper and Lower Sorbian are closely related to Polish and Czech. There is an active language community working on the preservation of these languages, namely the WITAJ-Sprachzentrum[3] (WITAJ Language Center) who also provided parts of the data used in this shared task. Previously, the WMT Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT (Weller-Di Marco and Fraser, 2022; Libovický and Fraser, 2021; Fraser, 2020) focused on both languages.

Ukrainian (`ukr`; `ukra1253`), spoken by approximately 40 million L1 speakers worldwide, is considered a mid-resource language in NLP with already several language-specific pre-training corpora (Chaplynskyi, 2023) and LLMs (Yukhymenko et al., 2025) available. With a significant number of Ukrainians currently living abroad, the demand for high-quality machine translation systems to support integration into new environments is greater than ever. Machine translation, complemented by cross-lingual knowledge transfer and robust question answering capabilities, can play a crucial role in addressing this need.

2.2 Task Description

Our main goal is to observe the synergy between the two different tasks in an LLM. Therefore, models are tested *jointly* on two tasks: **machine translation** and **multiple-choice question answering**. All participating systems must submit outputs for both tasks of a track.

For Machine Translation, we focus on the following translation directions:

- German to Upper Sorbian (de→hsb)
- German to Lower Sorbian (de→dsb)
- English to Ukrainian (en→uk)
- Czech to Ukrainian (cs→uk)

All these pairs focus on the more challenging—and needed—direction, from the higher-resourced language to the shared task languages.

For Question Answering, we use multiple-choice datasets from education and language certification. For Ukrainian, we evaluated on multiple-choice exam questions from the UNLP 2024 Shared Task on LLM Instruction-Tuning for Ukrainian, which is compiled from school graduation examinations on various subjects. For Upper and Lower Sorbian, we evaluated on language certificate exercises which follow the CEFR (Common European Framework of Reference for Languages) scheme.

2.3 Models and Restrictions

We set this shared task in a restricted context with limited resources: The base LLM is fixed to the Qwen 2.5 family (Qwen Team et al., 2025; Yang et al., 2024) and a maximum of 3B parameters. We chose a small model size in order to enable teams with fewer compute resources to participate, and limited models to one family so that results would be more readily comparable. The Qwen 2.5 model family was selected based on zero-shot performance on the Sorbian QA development sets.

Regarding data resources, we suggested training data and provided development sets for all languages (§3). However, the participants were not restricted to the provided datasets. Additional

[1] https://www2.statmt.org/wmt25/translation-task.html

[2] https://glottolog.org

[3] https://www.witaj-sprachzentrum.de

datasets were permitted under the condition that the used resources were open source.

2.4 Phases

The shared task was held in two phases: the development phase started with the release of training and development datasets (when available), and the test phase began when the respective test datasets were made available. The latter also featured a leaderboard on OCELoT where participants could submit their outputs to compare against other teams.

3 Datasets

In the following, we describe the datasets provided for the development and test phases.[4]

3.1 Sorbian Languages

We release the new data for both Sorbian languages under the CC BY-NC-SA licence.

MT For Machine Translation, the WITAJ-Sprachzentrum provided new monolingual and parallel datasets compared to the previous shared task editions for both Sorbian languages. The development dataset is a combination of both old and new sentence pairs. Table 1 lists the number of sentences or sentence pairs with German translations.

The test sets contain 4,000 sentences for each language. The sentence pairs are from the same domain as the training and development datasets.

	Upper Sorbian	Lower Sorbian
parallel train	187,270	171,964
parallel dev	4,000	4,000
parallel test	4,000	4,000
monolingual	47,758 (wiki) 1,071,723 (witaj)	120,501

Table 1: Number of sentences in the newly released mono- and bilingual (paired with German) data for both Upper and Lower Sorbian.

QA The WITAJ-Sprachzentrum organises language examinations for both Upper and Lower Sorbian, from the A1 to C1 levels, according to the CEFR scheme. We use a mix of questions from all five available levels (A1, A2, B1, B2, and C1, from beginner to advanced) for our Question Answering task—more specifically from the reading, grammar, and listening parts. For the listening questions, we rely on the reference transcription of the audio material to convert the exercises; this can lead to comparatively easier questions.

Each language level differs in terms of exercise formats, which can add another level of complexity to the task. While the beginner levels (e.g., A1) often have true-or-false questions on a short text, the advanced exercises (e.g., B2 or C1) typically consist of multiple-choice questions with longer texts and statements, sometimes with up to 16 possible answers.[5]

While the provided development set only contained questions from the A1 to the B2 levels, the test dataset additionally features questions from the C1 certification level. Table 2 shows the number of questions per difficulty level for both data splits.

split	lang	A1	A2	B1	B2	C1	total
dev	hsb	30	28	44	56	0	158
dev	dsb	30	28	44	56	0	158
test	hsb	30	28	44	56	52	210
test	dsb	29	28	44	56	48	205

Table 2: Number of questions in the Sorbian QA datasets per language level.

3.2 Ukrainian

The Ukrainian language track data was sourced from previous editions of MT and QA shared tasks, with the MT test set aligned with this year's General MT competition.

MT For the machine translation subtask, we focused on translation into Ukrainian from two languages: English→Ukrainian (en→uk) and Czech→Ukrainian (cs→uk).

The suggested development data were the combined WMT datasets from the 2022–2024 editions (Kocmi et al., 2022, 2023, 2024). Our test phase was aligned with the WMT 2025 General Machine Translation task for translation into Ukrainian. Parallel data statistics per language pair are shown in Table 3.

The primary difference in dataset sizes comes from the fact that most training and development sets were sentence-based, whereas the test set was

[4]Available at: https://github.com/TUM-NLP/llms-limited-resources2025

[5]More details are available in our shared task repository.

language pair	dev	test
cs→uk	6,263	230
en→uk	5,108	86

Table 3: Datasets statistics per language pair for the Ukrainian MT track.

designed at the paragraph level. This dissimilarity introduced additional challenges for both shared task participants and the evaluation process.

QA We utilised the dataset from the UNLP2024 shared task (Romanyshyn et al., 2024), focusing exclusively on multiple-choice questions. The original data splits were retained, ensuring balanced coverage of all topics across training, development, and test sets, as shown in Table 4.

split	Ukrainian history	Ukrainian lang. and lit.	total
train	910	1,540	2,450
dev	228	385	613
test	348	403	751

Table 4: Datasets statistics per splits and subjects for the Ukrainian QA track.

This dataset comprises machine-readable questions and answers from the Ukrainian External Independent Evaluation (transl. ZNO), the standardized examination required for university admission in Ukraine. It includes exam materials from 2006 to 2023, covering two subjects: History of Ukraine and Ukrainian Language and Literature.

3.3 Relative Difficulty of the Tracks

While we compiled the same task types for both tracks, there are different challenges for our selected languages.

For Ukrainian MT, natural parallel training data is already available, and a variety of both open-source and proprietary translation systems exist. Participants were therefore encouraged to leverage available resources to whatever extent they find appropriate. In contrast, Sorbian resources include parallel and monolingual sentences but are fewer in comparison. NLP tools, more generally, are not available for the two Sorbian languages.

For the QA task, the Sorbian dataset was derived from language certification exams across different levels. Given the languages' low-resource status, the focus is on evaluating whether models can adequately comprehend the language and answer typical document-understanding or grammatical questions. For Ukrainian, the challenge extends further: Models are tasked not only with answering general language questions but also with addressing deeper questions related to Ukrainian literature and history.

4 System Descriptions

Submissions are language-specific; participants could submit to one or more language tracks. However, participants had to submit both the QA and the MT outputs, which had to be generated by the same model.

In addition to our baseline outputs, three teams submitted to the Upper and Lower Sorbian tracks, and one team also submitted to the Ukrainian track. For the final evaluation, each team was asked to choose a primary submission and provide a short description of their approach. Based on these descriptions, we provide an overview of the participating systems here. Table 5 summarises the main characteristics of the three participating primary submissions. For more details, please refer to the respective system description papers.

Baseline (TUM Organisers) Our simple baseline prompts Qwen2.5-3B-Instruct (Qwen Team et al., 2025) with no fine-tuning. The prompts are zero-shot, and an example is shown in Appendix A. We implemented our tasks in the LLM Evaluation Harness framework (Gao et al., 2024) and provided this code to participants for reference.[6]

Team NRC (National Research Council Canada) (Larkin et al., 2025) The NRC submissions focused primarily on the machine translation (MT) component of the shared task. The team explored the impacts of training on Upper Sorbian and Lower Sorbian data separately and together. They also experimented with direct preference optimisation using pairs of correct and incorrect responses to the question answering (QA) set, with limited success. Their submitted systems are based on the Qwen2.5-1.5B-Instruct model, trained using supervised fine-tuning on full model weights using LLaMa-Factory and 4 GPUs (Tesla V100-SXM2-32GB).

[6] https://github.com/TUM-NLP/wmt25-lrsl-evaluation

			NRC	SDKM	TartuNLP
Tracks		hsb & dsb tracks	✓	✓	✓
		uk track	✗	✓	✗
System		Base Qwen model	1.5B	3B	3B
		LoRA	✗	✓	✗
		Quantised	✗	✗	✗
Data	Sorbian MT	Previous WMT MT data	✓	✗	✓
		External data for MT	✗	✓	✓
	MT	Backtranslation	✗	✓	✗
	QA	External data for QA	✗	✓	✗
Training		Joint hsb+dsb training	✓	✗	✓
		Instruction tuning	✗	✗	✓

Table 5: Summary of the three **primary** submissions.

Team SDKM (JGU Mainz) (Saadi et al., 2025) Team SDKM trained three separate Qwen2.5-3B-Instruct models for three languages: Lower Sorbian, Upper Sorbian, and Ukrainian. For Lower Sorbian, the team trained a Qwen2.5-3B-Instruct model by combining both machine translation (MT) and question answering (QA) data. Then, they fine-tuned a joint model on a combined MT and QA dataset. After this fine-tuning, they fine-tuned the model on all the provided QA data and 3k MT data for a second round of fine-tuning. This final fine-tuned model was used for both MT and QA tasks. For Upper Sorbian (hsb), the team followed the same overall approach as with Lower Sorbian. During the QA evaluation of dsb and hsb, they made multiple versions of the same MCQ question with different possible orders and averaged the likelihood of each option, then selected the option with the maximum likelihood. For Ukrainian, they also followed a similar approach, but for QA, they employed retrieval-augmented generation (RAG). The team will make all the data and code they used for pre-processing, training, and their trained model publicly available shortly after the deadline. They used the Qwen2.5-3B-Instruct model as all translation models, for semantic similarity calculation to incorporate retrieval augmented generation.

Team TartuNLP (TartuNLP) (Purason and Fishel, 2025) The TartuNLP system fine-tuned Qwen2.5-3B-Instruct (Qwen Team et al., 2025) on a mixture of monolingual, parallel, and instruction data to support both Lower and Upper Sorbian. The monolingual set included all sentence-level Sorbian data from current and past WMT Shared Tasks, Upper and Lower Sorbian Wikipedia articles (Foundation, 20250520 dump), and Upper and Lower Sorbian documents from Fineweb-2 (Penedo et al., 2025). Parallel data, sourced from WMT (current and past), was reformatted as chat-style instruction pairs, with four epochs of German-to-Sorbian and one epoch of Sorbian-to-German translations. Both monolingual and parallel Sorbian data were repeated four times. The instruction data was collected from Magpie (Xu et al., 2024), Aya (Singh et al., 2024), EuroBlocks (Martins et al., 2025), OpenAssistant (Köpf et al., 2023), and FLAN v2 (Longpre et al., 2023), covering multiple languages. All datasets were deduplicated and packed into 4096-token sequences, with loss applied only to assistant responses in instruction-formatted data. For the final submission, they used beam search with a beam size of 4 for machine translation and one-shot prompting with development set examples for question answering. The final model is published on HuggingFace.[7]

5 Primary Submission Results

5.1 Evaluation Methodology

We evaluate MT with chrF++ (Popović, 2015), computed using SacreBLEU (Post, 2018),[8] and QA with accuracy. ChrF++ ranked slightly higher than BLEU (Papineni et al., 2002) in the WMT 2024 Metrics Shared Task (Freitag et al., 2024). Although neural metrics such as COMET are gen-

[7] https://huggingface.co/tartuNLP/Qwen2.5-3B-Instruct-hsb-dsb

[8] SacreBLEU chrF++ signature: `nrefs:1 | case:mixed | eff:yes | nc:6 | nw:2 | space:no | version:2.5.1`.

erally known to better correlate with human judgements, we could not consider them as the main ranking criterion because they do not support the Sorbian languages. However, we do consider xCOMET (Guerreiro et al., 2024) for Ukrainian in Section 6.1.

Since our goal is to evaluate the *joint* performance of LLMs on both MT and QA, the ranking in the leaderboard takes into account the scores from all tasks of the track equally. For the final ranking, points are given according to the ranking of the submission in the MT and QA tasks, with the highest-ranked system obtaining the maximum number of points (4 for the Sorbian tracks, 2 for the Ukrainian track). In case of ties, we ranked according to the MT results.

5.2 Leaderboard Results

Tables 6, 7, and 8 show the results of the participating teams' primary submissions for the three tracks: Upper Sorbian, Lower Sorbian, and Ukrainian. The winning team per track and the best submission per task are in bold.

	de→hsb		hsb-QA		final
	chrF++	pts	acc.	pts	pts
TartuNLP	86.33	4	**58.10**	4	8
NRC	**87.20**	4	29.05	1	5
SDKM	75.73	2	55.24	3	5
baseline	13.88	1	42.86	2	3

Table 6: Results for the primary submissions for the **Upper Sorbian** track, ranked by number of points (pts).

	de→dsb		dsb-QA		final
	chrF++	pts	acc.	pts	pts
TartuNLP	78.20	4	**57.56**	4	8
NRC	**78.24**	4	32.20	1	5
SDKM	64.34	2	51.71	3	5
baseline	12.21	1	45.85	2	3

Table 7: Results for the primary submissions for the **Lower Sorbian** track, ranked by number of points (pts).

Upper and Lower Sorbian tracks For both Upper and Lower Sorbian tracks, TartuNLP was the overall winner with high results in both MT and QA. Looking at the tasks of translation and question answering separately, all systems outperformed the baseline for translation, while one system remained below the baseline for question answering. Indeed, if we focus on the translation results only, NRC obtained similar or better performance than TartuNLP, but it showed noticeably lower results for the QA task, since the team chose to focus on MT performance. As they reported, exclusively fine-tuning for MT affects the QA performance negatively, with lower accuracy than the baseline. This confirms our initial assumption and answers our first research question. On the other hand, the NRC submission relies on the smaller 1.5B model and manages to compete with the larger 3B model effectively. This is a promising result for low-resource MT.

For QA, the improvements remain more modest in comparison, with the accuracy increasing by 15 and 11 points, for Upper and Lower Sorbian, respectively. The lack of dedicated training data in the language and the variety of the exercises seem to be the main reasons preventing the models from reaching higher scores.

Ukrainian track For the Ukrainian track (Table 8), only one team (SDKM) participated, achieving results that slightly outperformed the baselines for both MT and QA. In MT, the gains were rather for the closely-related cs→uk pair than for the more distant en→uk pair. Despite accounting for differences in input style between the development and test sets, the results indicate that translating more complex, document-level content remains a substantial challenge for Ukrainian. In QA, the team also surpassed the baseline. For broader comparison, in the UNLP2024 shared task (Romanyshyn et al., 2024), the best-performing model based on Mistral-7B achieved an accuracy of 49, highlighting that smaller models still struggle to reach competitive performance on this task.

6 Deeper Analysis and Discussion

This section focuses on a few analyses of the results. For Ukrainian, to approximate a more advanced evaluation, we compared the participants and the baseline MT results with xCOMET (Guerreiro et al., 2024) (§6.1). For the Sorbian language tracks, we check the QA accuracy per language level (§6.2), contrast the translation performance against other MT approaches (§6.3.1), namely against the current Sorbian-German translator, and perform a manual annotation of translation outputs (§6.3.2).

	cs→uk		en→uk		uk-QA		final
	chrF++	points	chrF++	points	acc.	points	points
SDKM	**8.09**	2	**2.98**	2	**35.82**	2	6
baseline	3.48	1	0.34	1	31.16	1	3

Table 8: Results for the primary submissions for the **Ukrainian** track.

6.1 Evaluating Ukrainian MT with xCOMET

Since the xCOMET metric (Guerreiro et al., 2024) has demonstrated strong performance as a neural-based automatic MT evaluation method and supports Ukrainian, we employed it to gain a deeper insight into the Ukrainian MT results.

We tried both xCOMET-XL[9] and xCOMET-XXL[10] to estimate the difference in significance between the baseline and SDKM team results. The results from xCOMET models are presented in Table 9.

lang. pair	x-wins	y-wins	stat.	p-value
xCOMET-XL				
cs→uk	0.09	0.88	-1.80	0.07
en→uk	0.44	0.48	-0.13	0.89
xCOMET-XXL				
cs→uk	0.01	0.97	-3.17	0.00
en→uk	0.42	0.48	0.04	0.96

Table 9: Comparison of the xCOMET results for the baseline (x) and SDKM (y) **Ukrainian MT** submissions with t-test results.

For the en→uk pair, both XL and XXL models confirmed that the differences between systems were not statistically significant. In contrast, for the cs→uk pair, the XL model results were borderline with respect to the null hypothesis, while the XXL model clearly indicated a significant difference. These findings confirm that the SDKM team achieved better performance for cs→uk translation.

In addition, we conducted a qualitative analysis of the MT outputs at the sample level. A native Ukrainian speaker evaluated the translations for both fluency and adequacy. Representative examples from both models are included in the Appendix B.2. We observe that both models struggled with paragraph-level translation. In several cases, the baseline system failed to generate any output at all. The SDKM team's results, however, are particularly interesting: although their translations did not reproduce the full paragraphs, they captured the main content, resembling a blend of translation and summarisation. This suggests that, with more granular input pre-processing, the model could potentially produce more accurate translations.

[9] https://huggingface.co/Unbabel/XCOMET-XL

[10] https://huggingface.co/Unbabel/XCOMET-XXL

6.2 Detailed Sorbian QA results

model	A1	A2	B1	B2	C1
Upper Sorbian					
TartuNLP	**86.67**	**82.14**	**56.82**	37.50	**51.92**
NRC	50.00	32.14	22.73	19.64	30.77
SDKM	80.00	78.57	**56.82**	**41.07**	42.31
baseline	70.00	57.14	40.91	26.79	38.46
Lower Sorbian					
TartuNLP	**89.66**	71.43	**56.82**	**41.07**	**50.00**
NRC	55.17	42.86	18.18	26.79	31.25
SDKM	82.76	57.14	50.00	37.50	47.92
baseline	65.52	**75.00**	43.18	26.79	41.67

Table 10: Accuracy on the QA datasets per language level for **Upper and Lower Sorbian**.

Table 10 presents the details of the Sorbian QA results for each language level (A1 to C1). We observe that the accuracy drops overall with more difficult question levels for both Upper and Lower Sorbian, except for the B2 level. The lower score at the B2 level for all models compared to the technically more difficult C1 level could be explained by the question types. We recall here that since the questions come from an actual language certification, they are diverse in terms of question type and number of possible answers (cf. §3.1). This also means that we gave equal weight to comparatively more difficult and simpler questions in the main evaluation. If we choose a passing accuracy of 50, most submissions reach a B1 level approximately.

6.3 Detailed Sorbian MT Results

6.3.1 Comparing Sorbian MT

For both Upper and Lower Sorbian, we compare the performance of the submitted systems with existing MT-specific models and quantised versions of LLMs. We present two types of models: the current MT model from the WITAJ-Sprachzentrum (sotra) and a quantised version of the TartuNLP model.

sotra Sotra[11] is the Machine Translation platform developed by the WITAJ-Sprachzentrum since 2019. Dedicated models translate from and to four languages: Upper Sorbian, Lower Sorbian, German, and Czech (except the German-Czech pair), as of 2025. It is based on 800MB fairseq models (Ott et al., 2019), and 50MB quantised versions (INT8 with CTranslate2) have been used for the online version for notably faster outputs. We present the scores for both systems.

Quantised version of the TartuNLP submission Since the sotra website relies on the quantised version for faster inference, we also quantised the model submitted by the TartuNLP team in two different ways. More precisely, we consider a Q4_K_M and a Q8_0 quantised GGUF version of the model.

	de→hsb	de→dsb
NRC	**87.20**	**78.24**
TartuNLP	86.33	78.20
SDKM	75.73	64.34
baseline	13.88	12.21
sotra quantised	79.07	75.92
sotra unquantised	81.52	77.38
TartuNLP Q4_K_M	83.96	75.55
TartuNLP Q8_0	84.83	76.65

Table 11: Comparison of the chrF++ scores on the Upper and Lower Sorbian test dataset for different MT systems.

Results Table 11 presents the MT results on the same test dataset for sotra models and the quantised TartuNLP models. We first observe that the best MT submissions (NRC and TartuNLP) are better than the current Sorbian translator, sotra, for both languages and even with the unquantised version. The gap is larger for Upper Sorbian than Lower Sorbian. We note, however, that the sotra models are older and are thus trained with less data. They are also smaller with around 56M parameters.

Besides, as expected, more aggressive quantisation leads to worse performance. For instance, the Q4_K_M version of the TartuNLP model slightly underperformed compared to the current online sotra model in Lower Sorbian. For Upper Sorbian, the systems still performed better. The inference time is, however, still in favour of the sotra model.

6.3.2 Manual rank annotation of Sorbian MT

As more advanced and reliable automatic metrics are not available for both Sorbian languages, we also evaluate the machine translation outputs through manual rank annotation, despite the high human and time cost associated with it.

Annotation methodology For each Sorbian language, one native speaker ranked the translations from the four systems (including the baseline) for 60 sentences, thanks to the joint organisation with WITAJ-Sprachzentrum. To select the sentences, we first filter the test set (and translations) to avoid cases where two or more systems output the same or too similar sentences (especially for short sentences) or obtain extreme chrF++ scores (e.g., issues with the generation or perfect translation). Then, we randomly select 60 sentences and shuffle the translations before the manual annotation, to reduce bias from the order of the systems.

The two annotators were given the same instructions regarding the ranking. Ranks are assigned to the four translations of the sentence, from 1 for the best translation to 4 for the worst. If two translations are of similar quality, the same score rank can be given (e.g., 1, 2, 2, 4). The reference translation is also given for information purposes. Annotators can additionally put comments beside each machine translation.

Results Table 12 compares the MT system rankings produced according to our main metric (chrF++ score) and the human evaluation for both languages. Unsurprisingly, the baseline model is consistently ranked last by the human annotator for both Sorbian languages; the higher ranks achieved in Lower Sorbian are only due to a large number of ties (e.g., [1, 2, 2, 2]), which blur their poor absolute performance here. The best submitted system is, however, more difficult to conclude; as with the chrF++ score, the annotators also found the NRC and TartuNLP model outputs to be of higher and

[11] https://sotra.app

	rank	chrF++				human			
		NRC	TartuNLP	SDKM	base.	NRC	TartuNLP	SDKM	base.
Upper Sorbian	1st	28	22	10	0	23	26	11	0
	2nd	18	29	13	0	24	25	12	0
	3rd	14	9	37	0	13	9	37	0
	4th	0	0	0	60	0	0	0	60
Lower Sorbian	1st	34	22	4	0	45	37	22	1
	2nd	24	32	4	0	14	18	16	7
	3rd	2	6	52	0	1	5	22	15
	4th	0	0	0	60	0	0	0	37

Table 12: System rankings according to the chrF++ score and the human evaluation for 60 sentences. For instance, the NRC translations in Upper Sorbian were ranked first among the four systems for 28 sentences according to chrF++, while it was 23 times according to the human evaluation.

similar quality. Interestingly, for Upper Sorbian, the latter model seems to be better with the human evaluation.

We also count how often the rankings according to human annotations and chrF++ perfectly match. 27 system rankings (out of 60) are identical for Upper Sorbian and 3 for Lower Sorbian. This difference is due to the higher number of ties given to the systems in the manual annotation of Lower Sorbian.

Qualitatively, the Lower Sorbian annotation comments also showed that the output machine translations still remain unsatisfactory overall, even for the best-ranked system. We present selected examples for both languages in Appendix B.1.

7 Conclusion

The WMT 2025 Shared Task LLMs with Limited Resources for Slavic Languages was the first attempt to evaluate two tasks *jointly*, Machine Translation and Question Answering, and to assess how they impact each other. We focused on three Slavic languages: Upper and Lower Sorbian (paired with German for MT), and Ukrainian (paired with Czech and English). Submissions were constrained to open-source datasets and Qwen 2.5 models below 3B parameters for reproducibility.

Three teams participated. TartuNLP was the overall winner by jointly fine-tuning the model using Sorbian and instruction datasets. NRC won both Sorbian MT tasks with a smaller 1.5B model by focusing on the MT task only. SDKM submitted to all three tracks using additional external datasets as well as data augmentation with machine translation and won on the Ukrainian track. All submissions improved the Machine Translation quality over the baseline Qwen 2.5 3B model. We observe that only focusing on MT negatively affects QA performance, answering our first research question. However, improving the model capabilities on two different tasks remains possible even for small LLMs. This first shared task paves the way for an extension to other tasks and low-resource languages.

Ethics Statement

The shared task focused specifically on smaller models, limiting the submissions to LLMs below the milestone of 3B parameters. This lowers the computational barrier for participants, fostering accessibility and a smaller ecological footprint.

We also strove to have the results reproducible by ensuring that the teams only use both open-source models and datasets. Participants were encouraged to make their model public.

For the Sorbian tracks, the shared task relies on the partnership with the WITAJ-Sprachzentrum. They provided all new Sorbian datasets for both tasks. For the manual evaluation experiment of the Sorbian MT outputs, both annotators were contacted through the WITAJ-Sprachzentrum.

Limitations

The main limitations of this shared task come from the restricted resources. In-domain training data was scarce overall, if not completely lacking, as in the Sorbian QA task, compared to other high-resource languages. Hence, participants needed to resort to data augmentation or external data with machine translation to circumvent this constraint.

Besides, diverging data formatting might have added another layer of complexity, which is orthogonal to our research question. For Ukrainian MT, the sentence-level training data contrasted with the document-level input in the test set, and for Sorbian QA, the exercise variety in type and number of possible answers proved to be a challenge.

Finally, the shared task focused on track-based (i.e., language-specific) approaches for models and not a fully multilingual LLM for all three Slavic languages. The submissions to the Sorbian tracks showed that fine-tuning with both languages proved to be mutually beneficial.

Acknowledgements

This work has received funding from the European Research Council (ERC) under grant agreement No. 101113091 - Data4ML, an ERC Proof of Concept Grant.

We thank the UNLP 2024 Shared Task team: Roman Kyslyi, Mariana Romanyshyn, and Oleksiy Syvokon, for sharing the Ukrainian QA resources.

Moreover, we are grateful for our cooperation with WITAJ-Sprachzentrum for Upper and Lower Sorbian. Beyond the datasets they provided for the previous WMT shared tasks on MT for low-resource languages, this first edition relied on QA data. We thank Tomaš Šołta for providing the language certificate exercises in that regard. We also thank the annotator of Lower Sorbian MT outputs for his time and comments.

Finally, we thank the organisers of the WMT 2025 shared tasks. Firstly, we highly appreciate the help from Martin Popel, Tom Kocmi, Katia Artemova, and Mariya Shmatova for sharing test sets for the Ukrainian MT track with us. We are also especially grateful to Vilém Zouhar and Daniel Deutsch for their advice, Philipp Koehn for helping us with Softconf, and Roman Grundkiewicz for setting up the leaderboard on OCELoT.

References

Dmytro Chaplynskyi. 2023. Introducing UberText 2.0: A corpus of Modern Ukrainian at scale. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Wikimedia Foundation. `https://dumps.wikimedia.org/`.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations - democratizing large language model alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 47669–47681. Curran Associates, Inc.

Samuel Larkin, Chi-kiu Lo, and Rebecca Knowles. 2025. NRC Systems for the WMT2025-LRSL Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical report. *Preprint*, arXiv:2506.04079.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. Efficient continual pre-training of LLMs for low-resource languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlícek, Vinko Sabolcec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *Preprint*, arXiv:2506.20920.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Taido Purason and Mark Fishel. 2025. TartuNLP at WMT25 LLMs with Limited Resources for Slavic Languages Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Hossain Shaikh Saadi, Minh Duc Bui, Mario Sanz-Guerrero, and Katharina von der Wense. 2025. JGU Mainz's Submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. 2024. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia. ELRA and ICCL.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm: An efficient state-of-the-art ukrainian llm.

A Baseline Prompts

We show the examples of the MT and QA prompts used in the baseline. Below, we first show the MT prompt for Czech to Ukrainian. The other translation directions match this template, with the language names and ISO-3 codes substituted.

```
Translate the following Czech text to Ukrainian.
Put it in this format <ukr> Ukrainian translation </ukr>.
<cze> {{src_text}} </cze>
```

Below is the QA prompt. The possible answers for Ukrainian are А, Б, В, Г, and Д, so the output of the model is constrained to these 5 options in our evaluation of the baseline. Similarly, the Upper and Lower Sorbian answers are constrained to 1 to 16. While there are not always 16 possible options for each question, we did not observe the baseline model ever choosing a value outside the actual range given in the prompt.

```
{{context}}

Question:
{{question}}

Possible answers:
{{possible_answers}}

Answer:
```

B Illustrative MT Examples per Language Track

We provide several illustrative examples per language track to showcase major successes or problems of the submitted models.

B.1 Upper Sorbian and Lower Sorbian MT Examples

Tables 13 and 14 present two examples of both Upper and Lower Sorbian machine translation. We selected sentences for which notable phenomena are visible; they are, however, not representative of the overall system performance. Hence, we provide the chrF++ score and the translation rank from the human evaluation for information purposes.

The main issue with the baseline translations for both tracks is the language of the output sentence: it frequently features words from the German sentence (sometimes corrupted) or non-existing words in the Sorbian languages or German.

All three submissions improve largely on this issue. We observe a high similarity overall between the translations; if it was easy to identify the baseline output, annotators had to differentiate the other translations sometimes based on how well some nuances were captured.

We also present an output from the SDKM submission in Lower Sorbian (first sentence in Table 14) where non-Latin script characters suddenly appeared. This happened only once in the annotated sentences (i.e., out of 60).

B.2 Ukrainian MT Track Examples

		chrF++	human
de source	Dissoziieren oder sich häufig benebeln und abschalten sind keine günstigen Strategien zur Bewältigung der Situation.		
en translation	*Dissociating or frequently numbing yourself and switching off are not effective strategies for coping with the situation.*		
hsb reference	Disociěrowanje abo so husto zamućić a wotšaltować njejsu přihódne strategije k zmištrowanju situacije.		
baseline	Diszozierun ór sèfórfi ór abefeln ëd abschaltun sàs nesiónn sàs strategyn nòg dżewen.	12.27	4
NRC	Disociěrować abo so husto zamućić a wotšaltować njejsu žane dobre strategije k zmištrowanju situacije.	82.97	2
SDKM	Dissociěrować abo so husto pohłušić a wotpinać njejstej žanej spomóžnej strategiji k zmištrowanju situacije.	56.74	3
TartuNLP	Disociěrować abo so husto zamućić a wotšaltować njejsu žane přihódne strategije k zmištrowanju situacije.	90.06	1
de source	Und paradoxerweise verhalf ihm eben dieser Ultradogmatismus durch alle Spaltungen, Intrigen, Säuberungen hindurch zum Durchbruch.		
en translation	*And paradoxically, it was precisely this ultra-dogmatism that helped him achieve his breakthrough through all the divisions, intrigues, and purges.*		
hsb reference	A paradoksnje dopomha jemu runje tutón ultradogmatizm přez wšě rozpačenja, intrigi, čisćenja k předobyću.		
baseline	Ie ultradogmatizmus verhal iem dëm dazulieb, dëm derselbe dazulieb, dëm derselbe durch alle spaltungen, intrigen, säuberungen hiddur kom.	21.95	4
NRC	A paradoksnje dopomha jemu runje tutón ultradogmatizm přez wšě pačenja, intrigi, wučisćenja přez předobyće.	81.23	3
SDKM	A paradoksnje dopomha jemu runje tutón ultradogmatizm přez wšě pačenja, intrigi, čisćenja k předobyću.	93.54	1
TartuNLP	A na paradoksne wašnje dopomha jemu runje tutón ultradogmatizm přez wšě pačenja, intrigi, čisćenja k předobyću.	87.57	2

Table 13: Examples from the **de→hsb** MT track submissions. We also present the chrF++ scores alongside the human rank annotation.

		chrF++	human
de source	Er nannte als gutes Beispiel die Talsperre Versetal in Nordrhein-Westfalen.		
en translation	*He cited the Versetal dam in North Rhine-Westphalia as a good example.*		
dsb reference	Wón jo pomjenił ako dobry pśikład gaśeński jazor Versetal w Nordrhein-Westfalskej.		
baseline	Er nannte als gutes Beispiel die Talsperre Versetal in Nordrhein-Westfalen.	28.07	4
NRC	Wón jo pomjenił ako dobry pśikład rěcnu zawěru Versetal w Nordrhein-Westfalskej.	78.59	1
SDKM	Wón jo pomjenił ako dobre pśikłady rěcnu zawěru Wortetzer峡在 Nordrhein-Westfalskej.	57.98	2
TartuNLP	Wón jo pomjenił ako dobre pśikłady rěcnu zawěru Versetal w Nordrhein-Westfalskej.	70.57	2
de source	Ob sie wohl jener Mann gesandt hat, dachte Matej und stapfte in den unbekannten Wald.		
en translation	*Matej wondered whether that man had sent him and trudged into the unfamiliar forest.*		
dsb reference	Lěc jo jich ten muski pósłał, jo Matej pómyslił a stupał do njeznateje góle.		
baseline	Matej pohajowat pohadzowat, że gołomany měnka go do niekowiednega węgla wysłał, i chodzil w ten węglu.	19.83	4
NRC	Lěc drje jo jich ten muski pósłał, jo myslił Matej a jo stupał do njeznateje góle.	78.26	1
SDKM	Abo jo wóna togo muskego pósłała, jo se Matej pódał do njeznateje góle.	45.61	3
TartuNLP	Lěc drje jo jich ten muski wupósłał, jo myslił Matej a stupał do njeznatego lěsa.	63.81	2

Table 14: Examples from the **de→dsb** MT track submissions. We also present the chrF++ scores alongside the human rank annotation.

Original	Baseline	SDKM
Vypravěčem spletitého příběhu, točícího se kolem ukradeného velkého diamantu, je neúspěšný boxerský promotér zvaný Turek, který se k divokému kolotoči událostí připletl se svým komplicem Tommym vlastně náhodou. Ukradený diamant má předat kurýr Franky Čtyřprsťák americkému šéfovi Avimu. O drahokam však mají zájem další gangsteři a zlodějíčci, mezi nimi především bezohledný ruský zabiják Boris Břitva, černošští provozovatelé zastavárny Vincent a Solomon a nakonec i nemilosrdný gangster a organizátor nelegálních boxerských zápasů Kruťák...	'[invalid]'	Авторська гангстерська комедія Ґю Річі, з чеською назвою Підфус (1998), базується на попередньому фільмі того ж жанру, Бейсбол, викиньте гроші (1994).
Triptany = specifická léčba akutní migrény. Neuroložka Markéta Tučková a moderní léčebné postupy • mujRozhlas V tuto chvíli vítáme ve studiu ČRo Hradec Králové paní doktorku Markétu Tučkovou z oddělení neurologie Oblastní nemocnice v Náchodě. Dobré dopoledne. Dobrý den. Naše dnešní téma jsou bolesti hlavy, migréna, léčba moderní metody. To všechno bychom dnes chtěli stihnout, i když je to samozřejmě moc. Na úvod tedy, paní doktorko, mohli bychom si rozdělit, udělat představu, jaké jsou dnes možnosti, jak pomoct lidem s migrénou v tom smyslu, jaké léky existují?...	Триптани - це специфічна лікування акутної мигрені. Неврологка Маркета Тучкова і сучасні лікувальні пісокупи • міюРозхлід	Так, тепер профілактика, що є ще однією категорією, є насправді профілактичними препаратами, які не призначені для лікування мігрені, але вони приймаються для зменшення ризику розвитку мігрені, наприклад антигепатичними препаратами, які є моноклональними антитілами проти рецепторів глютену, і це також біологічна терапія мігрені.

Table 15: Examples from **cs→uk** MT track submissions. Both systems have major problems of not being able to capture full paragraph content.

Original	Baseline	SDKM
Start to the season she wanted off and so likewise 83 also off so that brought out another safety car. Christopher parrott involved in that but then great fights for third place especially as costello took the lead. Toth involved in a battle for second, third and fourth. A couple more offs that brought out another safety car. Some racing incidents. Hodson great restart finally to go into second place and that's what nailed that for him but no one was gonna stop michael costello who wins.	Почати сезон вона хотіла і тоді також 83 і офс так щоб піднімнувши ще один сейчар. Кристіпер Парротт був у цьому, але потім великий бой за третье місце особливо як Коостелло викризував перед. Товщина бойу за друге, третье і четверте. Другі пара офсі щоб піднімнувши ще один сейчар. Кількa ринкових подій. Ходсон великий починок навіть до постачання другого місця і це завдає йому це, але жодного не зупинить Майкела Коостелло який перемагає.	Починаючи з сезону, вона хотіла вийти, тож так само і 83, вони обоє були вийняті, що привело до чергового запуску безпеки. Крістофер Парротт був причетний до цього, але потім великі боротьби за третє, четверте і п'яте місце. Дещо аварійних інцидентів. Ходсон був у грі за друге місце, але ніхто не зупинить Майкла Костелло, який виграє.
Didn't expect that if i'm being a hundred percent honest. I didn't expect that i wasn't gonna make a video initially but i put out a tweet saying yay, and there was already a lot of confusion, which is understandable there is so much confusion around what is happening in america right now with the tariffs and there's a couple of points that i want to get to with that ...	'[invalid]'	'[invalid]'

Table 16: Examples from **en→uk** MT track submissions. Both systems have major problems of not being able to capture full paragraph content.

Findings of the First Shared Task for Creole Language Machine Translation at WMT25

Nathaniel R. Robinson[1], Claire Bizon Monroc[2], Rasul Dent[3], Stefan Watson[4], Kenton Murray[1], Raj Dabre[5], Andre Coy[4], Heather Lent[6]

[1]Johns Hopkins University, USA; [2]Mines Paris PSL, France;
[3]Inria Paris, France; [4]University of the West Indies at Mona, Jamaica;
[5]IIT Madras, India; [6]Aalborg University, Denmark

Abstract

Efforts towards better machine translation (MT) for Creole languages have historically been isolated, due to Creole languages' geographic and linguistic diversity. However, most speakers of Creole languages stand to benefit from improved MT for low-resource languages. To galvanize collaboration for Creole MT across the NLP community, we introduce the First Shared Task for Creole Language Machine Translation at WMT25. This Shared Task consists of two systems tracks and one data track, for which we received submissions from five participating teams. Participants experimented with a wide variety of systems and development techniques. Our evaluation campaign gave rise to improvements in MT performance in several languages, and particularly large improvements in new testing genres, though some participants found that reusing subsets of pretraining data for specialized post-training did not yield significant improvements. Our campaign also yielded new test sets for Mauritian Creole and a vast expansion of public training data for two Creole languages of Latin America.

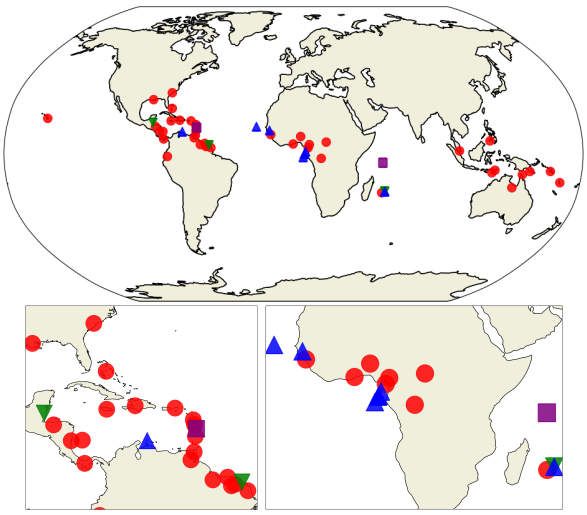


Figure 1: Creole languages included in the shared task, plotted geographically. Purple squares are languages for which we received *constrained* system submissions. Blue up-triangles are those with *unconstrained* system submissions, and green down-triangles are for *data* submissions. Red circles are for remaining Creoles for which we supported submissions but received none.

1 Introduction

Insufficient training data remains a pronounced barrier for creating natural language processing (NLP) systems that cater to lower-resourced languages. This is particularly pronounced for the task of machine translation (MT), due to the importance of aligned bitexts. For Creole languages, a geographically and linguistically diverse group (see Figure 1), the lack of training data brings some challenges common to other low-resource scenarios, but also offers unique opportunities, due to the role of contact in shaping them.

Many Creole-speaking communities have expressed interest in having MT support for their language (Lent et al., 2022). However, efforts to create NLP systems for them have largely been fragmented. This is in part due to the truly expansive scope of the term "Creole". By definition, a Creole language exhibits linguistic influence from an amalgamation of languages, typically both high- and low-resource (Kouwenberg and Singler, 2009). Naturally, the Creole languages of Africa (*e.g.* Nigerian Pidgin and Sango) are often viewed with different historical and cultural lenses than those of the Caribbean (*e.g.* Haitian Creole and Papiamento) or Pacific (*e.g.* Bislama and Tok Pisin).

However, this geographic and cultural fragmentation misses some of the notable commonalities among Creole languages. Many of the Creole languages of Africa and the Caribbean have a shared linguistic history: most of the languages in both groups emerged from European colonialism and slavery in Africa, and languages in both groups are strongly influenced by European and African

Proceedings of the Tenth Conference on Machine Translation, pages 520–531
November 8-9, 2025

(typically Niger-Congo) languages in similar ways (Gilman, 1986; Robinson et al., 2024). While Pacific Creole languages are not connected to Africa, they are also influenced by largely the same European colonial languages (*e.g.*, English, French, Portuguese, Dutch, *etc.*), and many show comparable and sometimes even more pronounced grammatical influence from Oceanic languages (Keesing, 1988). Perhaps even more important than their linguistic commonalities, Creole languages across the globe are subject to similar stigmas and play similar linguistic roles in relation to high-resource languages (DeGraff, 2003). Hence, speakers of Creole languages everywhere may have shared experiences in connection to their language and could benefit similarly from improved language technologies.

Only recent publications have addressed the data scarcity of large numbers of Creole languages, a necessary hurdle to create systems that truly serve speakers. (See Section 2.) These recent efforts have been ground-breaking (Lent et al., 2024; Robinson et al., 2024). However, their large scope allowed only for treatment of Creole languages as a conglomerate without particular attention to specific language communities and their distinct needs. To build technologies for the wide array of Creole language speakers, participation from a variety of communities is needed. In an effort to foster collaboration in NLP across these communities, we hold the first Shared Task for Creole Language MT held at WMT25.[1] We detail prior publications that inspired the need for this shared task in Section 2, and we overview the task organization in Section 3. The shared task consists of two subtasks: a call for **System Submissions**—itself consisting of two tracks, namely **Track 1** (Constrained), **Track 2** (Unconstrained)—and a call for **Dataset Submissions**. Five teams participated. Submissions and evaluation results are discussed in Sections 4 and 5, respectively.

2 Related Work

Creole languages are a product of intense linguistic contact. They usually draw most of their vocabulary from a single source language that is often referred to as the Creole language's lexifier.[2] Typologically, Creole languages tend to be more isolating than these lexifier relatives. However, attempts to define them as distinct typological class (*e.g.* Bakker et al., 2011) are contested (*e.g.* Fon Sing, 2017). In light of the typological debate, some (*e.g.* Mufwene, 2001) prefer to define Creoles by their history, which is closely associated with the long-term movement of people speaking mutually unintelligible languages.

Today, languages recognized as Creoles are spoken by over 180 million people (Lent et al., 2021). For some speakers, a Creole language is their mother tongue, and these speakers may be monolingual; for others, a Creole language can serve as a lingua franca within the broader, multilingual community. In surveying Creole language speakers, Lent et al. (2022) find that they highly desired MT support for their languages. Bird (2022) further highlights the opportunity for Creole language technology to serve as a bridge between higher and lower-resourced communities. He emphasizes that in contexts where Creoles act as a lingua franca, it may be more viable to develop language technologies for them, rather than going straight to even lower-resourced highly localized languages.

Despite the demand and utility of MT for Creole languages, they occupy a small portion of works in the broader MT community. When they are included, works tend to focus on individual languages. For example, the 2010 earthquake in Haiti prompted the rapid development of MT systems for the French-related Haitian language (Lewis, 2010). Early work by Dabre et al. (2014); Dabre and Sukhoo (2022) on Mauritian Creole acknowledged this trend, though they too developed technologies for a single language. Similarly, MT systems for West African Pidgin (Ogueji and Ahia, 2019) are highly relevant for other English-related Creole languages, like Guyanese Creolese (Clarke et al., 2024), but it has been difficult to coordinate multi-continental efforts.

The first effort towards a joint, large scale MT project for Creoles was proposed in CREOLEVAL (Lent et al., 2024) with the CREOLEM2M model, covering N-way translation between English and 26 Creoles. Building on this work, Robinson et al. (2024) contributed KREYÒL-MT for 41 Creole languages, including the first public, extensively multilingual MT dataset for Creole languages, and focusing on those spoken by the African diaspora. The resulting models have set the current state-of-the-art for Creole language MT.

Even with these advances, efforts to build Creole

[1] https://www2.statmt.org/wmt25/creole-mt.html

[2] This nomenclature is contested, however, as some academics argue it advances the misconception that Creole languages are fundamentally different from other languages (DeGraff, 2003).

technologies are not strongly unified. For instance, while Robinson et al. (2024) focus on Africa and the Americas, the potential relevance of these languages for MT efforts for Pacific Creoles (*e.g.*, Chavacano) (Vicente et al., 2024) remains an open question. Perhaps most pertinently, because both CREOLEVAL and KREYÒL-MT advanced developments for Creole languages as a conglomerate, focus on specific languages and their distinct needs was outside the scope of efforts. In organizing this shared task we hope to foster the beginnings of broader collaboration in Creole NLP. This way, members of specific Creole language communities may have a place to develop their own communities' desired technologies.

3 Shared Task Overview

3.1 Call for Systems Submissions

We solicited MT systems for translation between any number of Creole languages and English or French, for which we have paired data. (We also welcomed submissions for other pairs of a Creole and non-Creole language each, asking participants to justify how translation between such languages would be relevant to the affected community.)

The official train and dev sets provided to participants were the public KREYÒL-MT train and dev splits available on HuggingFace.[3] Participants were not permitted to use the data designated as *test* data from this dataset. We also provided, to any participants who requested them, additional training bitexts with English translations of Haitian, Papiamento, and Sango text from the Church of Jesus Christ of Latter-day Saints, still pending release on LDC.[4] We solicited submissions of systems in both **constrained** and **unconstrained** tracks. To construct our official evaluation sets for both tracks, we selected a random seed (kept private) and shuffled the public KREYÒL-MT test sets.

Constrained Track The purpose of this track was to explore better ways to model Creole MT with limited resources and allow researchers to explore smarter configurations than the simple ones that were used in past Creole language MT models. For instance, Robinson et al. (2024) focused primarily on their presentation of the KREYÒL-MT dataset without justifying all engineering choices used to train the KREYÒL-MT model. Thus, while their model was trained on 41 Creole languages at once, it was not clear whether this model was trained *optimally* for all 41 languages given the resources available. To have a track which is directly comparable to the original KREYÒL-MT model, therefore, we only accepted submissions in one of the 40 Creole languages included in the KREYÒL-MT set, with translation into or out of English and/or French only. The baseline model for this track was the `kreyol-mt-pubtrain` model available on HuggingFace,[5] which is the model trained on the public KREYÒL-MT dataset (Robinson et al., 2024). We note that this model was not trained on any data that was not made available to the participants. Moreover, participants were permitted to use this model however they wished in the constrained track, including as an initialization for fine-tuning, but they were not be permitted to use any other pre-trained models.

Unconstrained Track The purpose of this track was simply to encourage creation of state-of-the-art Creole language MT systems. In the unconstrained track, teams were allowed to use data from any source, and leverage any pre-trained models or LLMs. As the relaxed constraints in this track allowed participants to develop systems for any Creole language, we used two baseline models: `kreyol-mt-pubtrain` (the same as the constrained track)[6] and CREOLEM2M (Lent et al., 2024) (available on HuggingFace,[7] which covers a number of Creole languages not supported by the former[8]). That said, this track allowed for submission of any of the Creole languages supported by either of the baseline models; with translation into/out of English and/or French permitted for languages supported by KREYÒL-MT, and translation into/out of English only permitted for the languages supported by CREOLEM2M, accordingly. We used the same evaluation sets as in the constrained track for any languages supported by KREYÒL-MT, and were prepared to furnish additional eval sets for any other Creole languages.

[3] https://huggingface.co/datasets/jhu-clsp/kreyol-mt

[4] Further details: https://huggingface.co/datasets/jhu-clsp/kreyol-mt/blob/main/README.md

[5] https://huggingface.co/jhu-clsp/kreyol-mt-pubtrain

[6] Due to a miscommunication on our part, the EHOW team used `kreyol-mt`, the KREYÒL-MT model trained on both public and private data, as their baseline model throughout their experiments. Hence we used `kreyol-mt` as the baseline for their submissions.

[7] https://huggingface.co/AAU-NLP/CreoleVal-CreoleM2M

[8] https://github.com/hclent/CreoleVal/

For this unconstrained track, we were also prepared to accommodate submissions for Creole languages not supported by either baseline model, in which case we would require the participants to submit an evaluation set of their own creation, and we would employ the baseline models in a zero-shot setting to calculate baseline scores. However, this circumstance did not arise.

3.2 Call for Data Submissions

We also solicited contributions to Creole language MT training and evaluation sets. We requested data submissions to be in bitext formats with translations into *any* other language – not only English or French. (Again, we stipulated that submissions should justify why the non-Creole language would be relevant for the Creole language-speaking community).

We thus established the following requirements for any submitted datasets:

- Participants must show that 100% of translations were either translated or post-edited by competent native or proficient speakers of the source and target languages.
- Participants must prepare a data card[9] with each submitted dataset.
- Participants must be able to show that one of the languages in each submitted bitext is considered a Creole language, by citing adequate academic sources or other sufficiently convincing means.
- If submitting training data, we strongly encouraged participants to also develop an accompanying MT system and evaluate this system on a test set (either a test set of their own creation, which must be submitted along with the training data, or a previously published test set). Ideally, participants should show significant ($p < 0.05$) improvements in chrF++ (Popović, 2017) over the previous state-of-the-art open-source MT system for the given language pair. (To do this they must identify the previous SOTA model and make a compelling case for why it would be considered SOTA.) We committed to provide software to assist meeting this requirement as needed. If participants were not able to meet this requirement, we required that they provide other convincing evidence of the utility of their training set.
- If submitting a test set, participants must use it to evaluate performance of an MT model and provide compelling evidence that the model's performance on the test set aligns with conventional wisdom regarding the model's performance in the translation direction.

[9] See https://oldi.org/guidelines

3.3 Support for participants

During the course of the shared task, we aimed to support participants wherever possible. To this end, we provided tutoring, paper workshopping, and a dataset for manual translation (i.e. FLORES-200 English (NLLB Team et al., 2022)).

4 Shared Task Submissions

4.1 Constrained system submissions

We received two submissions for the constrained track, in which participants were only permitted to use provided training data: **KREY-ALL** (Ayasi, 2025) for Seychellois Creole translation into English, and **LUDOVIC MOMPELAT** (Mompelat, 2025) for translation between Martinican Creole and French.

4.1.1 KREY-ALL

KREY-ALL investigated joint training on typologically related Creole languages. They focused on Seychellois Creole (`crs`), and selected four additional languages known to be structurally similar, due to both shared francophone vocabulary and historical migration patterns: Mauritian Creole (`mfe`), French Guianese Creole (`gcr`), Louisiana Creole (`lou`) and Réunion Creole (`rcf`) (Papen, 1978).

Translation data from these languages were used in conjunction with Seychellois Creole data, with two tagging strategies: "All Kreyol" where all languages used the same language tag tag (`crs`), and "Specialized" where each language used its own tag. For each strategy, both full and partial (last 4-6 layers) fine-tuning were compared.

KREY-ALL found that using the same language tag for all languages and fine-tuning all model parameters were most effective. They also found that, although Mauritian Creole data was an order of magnitude more plentiful than Seychellois data, upsampling the Seychellois segments by more than 5x relative to the other languages was ineffective. Analysis of the model's embeddings revealed a distinct cluster for each language, with Seychellois having close proximity and overlap with Mauritian. However, there was no discernible Indian Ocean group, as Mauritan also overlapped with Louisiana

Creole, while Seychellois showed commonalities with Guianese. The other Indian Ocean language, Réunion Creole, did not overlap with any of the four.

4.1.2 LUDOVIC MOMPELAT

LUDOVIC MOMPELAT (LM) submitted two MT systems: from Martinican Creole (mart1259) to French and vice-versa. Their approach was to fine-tune kreyol-mt-pubtrain using LoRA (Hu et al., 2021). LM experimented with a different train/dev data ratio (70/30 vs. 90/10) and explored a few values of LoRA rank and scaling factor. Their final system for into-French MT used weighted BLEU scores for a curriculum sampling training set-up with difficulty ranking. Models for both translation directions were trained with label smoothing and a weighted BLEU criterion for checkpoint selection, and both employed a newly trained tokenizer with new language tags.

4.2 Unconstrained system submissions

Two teams submitted to the unconstrained track, in which participants were allowed to use external data and pre-trained models: **EDINHELSOW** (EHOW) (Rowe et al., 2025) and **KOZKREOLMRU** (Rajcoomar, 2025).

4.2.1 EDINHELSOW

EHOW submitted systems that translated between English and seven lusophone Creole languages: Angolar (aoa), Annobonese (fab), Guinea-Bissau Creole (pov), Kabuverdianu (kea), Papiamento (pap), Principense (pre) and Sãotomense (cri). The team conducted an incredibly thorough analysis of numerous techniques for enhancing Creole MT, and even leveraged the linguistic relationship between Portuguese and related Creoles.

Notably, EHOW collected additional parallel and monolingual data to supplement the provided data sources. These additional parallel data came from various sources: online Bible translations (pap and pov); the Jehovah's Witnesses' Watchtower magazine (kea, pap, and pov); text sourced from an online educational sentence generator (pap); and gloss text from a dictionary (pov). They used the monolingual corpora to create synthetic parallel data generated through back-translation, using the kreyol-mt model. For some of their experiments, the English sentences from the pap, kea, pov and cri bitexts were also forward-translated for Sequence-Level Distillation (Kim and Rush, 2016) of kreyol-mt. Again for some experiments, training data were further augmented using 112k high-quality English-Portuguese sentences, extracted from the Tatoeba Translation Challenge203 Dataset (Tiedemann, 2020).

Using these curated data sources, the EHOW team fine-tuned various pretrained multilingual base models: two sizes of NLLB (NLLB Team et al., 2022), three configurations of mBART (Tang et al., 2020), and kreyol-mt. EHOW experimented with inclusion of the Portuguese and distillation data mentioned in the previous paragraph, as well as with initializing language token embeddings with the Portuguese token embedding, to explore 14 combinations of training practice. They then merged six combinations of the resulting models to produce final systems. EHOW's *primary* submission for each language pair was the overall best performing merged model for each generalized direction (XX→eng or eng→XX), while *contrastive1* submissions were the best trained model (merged or otherwise) for each language pair. They found that post-editing of system outputs using LLMs and bilingual lexicons was typically not helpful, but they submitted some systems that incorporated this practice as *contrastive2*.

4.2.2 KOZKREOLMRU systems

KOZKREOLMRU took a unique three-step approach to translation between Mauritian Creole and English. Their first step was continuous pretraining of Llama 3.1-8B over 500k monolingual Mauritian Creole tokens (18k lines) sourced from (Dabre and Sukhoo, 2022), with an additional 100k monolingual tokens each of English and French data. Step two was then fine-tuning the model for MT. This was done on $\tilde{4}$0k lines of bitext, sourced again from Dabre and Sukhoo (2022); Robinson et al. (2024); 4.9k lines of synthetic data from prompting a Claude model with text from MMLU (Hendrycks et al.); and 300 lines of bitext from community translation of English Claude outputs. The final step was parameter-efficient fine-tuning (PEFT) via LoRA (Hu et al., 2021) over newly contributed Mauritian Creole translations of FLORES-200 (NLLB Team et al., 2022). In some eperiments, only the dev set translations were used for PEFT, while the devtest set was reserved for testing. In others, KOZKREOLMRU used all FLORES-200 data for PEFT and evaluated on a newly created bitext from the LALIT newspaper. (See Section 4.3 for details of these datasets.) The

KozKreolMRU team performed ablations, to isolate the effects of monolingual continuous pretraining, vanilla fine-tuning, and PEFT.

4.3 Data submissions

Two of our participating teams submitted datasets: **KozKreolMRU** (Rajcoomar, 2025) submitted two dev/test sets for Mauritian Creole↔English MT; and **JHU** (Robinson, 2025) submitted train, dev, and test sets for Belizean Kriol↔English and French Guianese Creole↔French MT. See data cards for these submissions on GitHub.[10]

4.3.1 KozKreolMRU data

KozKreolMRU submitted two dev/test sets to evaluate translation between Mauritian Creole and English. The first consists of Mauritian Creole translations of FLORES-200 (NLLB Team et al., 2022), containing 997 `dev` lines and 1012 `devtest` lines. This dataset is particularly useful because (1) it is automatically aligned with the other language sets contained in FLORES-200 and (2) FLORES is a common benchmark to judge MT model proficiency. The second dataset is a small test set consisting of 102 sentence pairs sourced from the LALIT newspaper.

4.3.2 JHU

JHU submitted three datasets for two language pairs. The first consists of 5.5k lines of Belizean Kriol and English translations from a Belizean textbook, both automatically and manually aligned after document processing. The second dataset comes from an online Bible translation and pairs 879 French Guianese Creole lines with French translations. The third dataset consists of 792 sentences from French Guianese Creole fables, aligned with French translations (mostly manually after web-scraping the raw data). All of these datasets were divided into train, dev, and test splits with a 90-10-10 ratio. Together they increase the amount of publicly available bitext by 2,300% for Belizean Kriol↔English and 370% for French Guianese Creole↔French. JHU demonstrated improvements ranging from +3.2 chrF++ to +33.3 chrF++ on the submitted test sets via fine-tuning `kreyol-mt-pubtrain` on the submitted train sets.

5 Evaluation Results

We designed a shared evaluation process for the two system tracks. All language pairs in the received submissions were supported by Kreyòl-MT, so only our shuffled Kreyòl-MT public test sets were used for official evaluation.[11]

For every language pair and direction, we computed the chrF++ (Popović, 2017) and BLEU (Papineni et al., 2002) scores of the submission using the default parameters of the `sacrebleu` library. We compared these results to the baseline scores of `kreyol-mt-pubtrain`, `kreyol-mt`, and CreoleM2M as stipulated in Section 3.1. Note that no two teams submitted a system for the same language pair, and therefore scores could only be compared with baselines. Results for the constrained track are reported in Table 1, and for the unconstrained track in Table 2. We discuss the results for each track separately below.

5.1 Constrained system results

direction	Team	Model	chrF++	BLEU
		creole → XX		
crs→eng	**Krey-All**	contrastive2	59.0	34.5
	Krey-All	contrastive1	58.9	34.2
	baseline	`kreyol-mt-pubtrain`	57.7	33.8
	Krey-All	primary	58.4	33.7
mart1259 →fra	*baseline*	`kreyol-mt-pubtrain`	50.4	28.3
	LM	primary	49.1	25.3
		XX → creole		
fra→ mart1259	*baseline*	`kreyol-mt-pubtrain`	48.7	26.5
	LM	primary	48.7	25.8

Table 1: Results of primary, *contrastive1*, and *contrastive2* submissions to the constrained track. Systems are ordered by BLEU score.

ChrF++ and BLEU scores are reported in Table 1 for both teams, alongside the `kreyol-mt-pubtrain` baseline. The LM systems score on par with or slightly below the baseline, while *constrative1* and *constrative2* submissions of Krey-All show marginal improvements.

The *primary* **Krey-All** submission corresponds to approaches with data all merged under the Seychellois Creole language tag (`crs`); *contrastive1* represents use of language-specific tags, and *constrative2* indicates partial parameter freezing (with the last 4 encoders, all decoder layers and

[10] https://github.com/n8rob/creolemt_wmt25

[11] This led to complications with the unconstrained systems track. Because Robinson et al. (2024) originally shuffled and split their private and public test sets independently, the `kreyol-mt` model trained on the private set has been contaminated with some of the segments in the public Kreyòl-MT test sets we used to construct our eval sets. Yet as a publicly available model, `kreyol-mt` was permitted for unconstrained submissions. This is what prompted our reevaluation of EHOW submissions; see Section 5.2.1.

the shared embeddings fixed during training). The best model for crs→eng was *constrative2*, however all four submitted systems score within one BLEU point of each other. This, combined with the observation that KREY-ALL came up with a different system ranking (one in which *constrative2* performed worst of all, (Ayasi, 2025)) by using a different shuffle of the same test set, suggests that score differences are not significant.

Recall from Section 4.1.2 that **LM**'s into-French system employed curriculum learning with BLEU-based difficulty ranking. The system for opposite translation direction did not employ this, but used different train/dev split (70-30 instead of 90-10). In our official evaluation, both submissions under-performed the baseline, albeit by less than 2.0 chrF++. LM's own reporting a slight improvement over the baseline for `mart1259→fra` on a shuffling of the same test set, again gives the impression that score differences are not significant.

Both LM and KREY-ALL fine-tuned the `kreyol-mt-pubtrain` model for a single language pair. These were valuable experiments because such attempts had not been made previously. The original work of (Robinson et al., 2024) focused primarily on the KREYÒL-MT dataset and did not include extensive experiments regarding how to engineer optimal MT systems from the available data. One question left open by their work was whether, if in the absence of additional data or models, significant improvements could be achieved by post-training on a select subset of training data. Results from these two studies (Ayasi, 2025; Mompelat, 2025) now indicate no such significant improvements, suggesting that either very different methods would be needed to improve performance, or that researchers may find more promise in developing new datasets and using external models (rather than post-training on subsets of `kreyol-mt-pubtrain`'s own data).

5.2 Unconstrained system results

ChrF++ and BLEU scores are reported in Table 2 for both teams, alongside baselines.[12] The **KOZKREOLMRU** systems are unique in that they are not based on any system that was trained on the full KREYÒL-MT dataset. Both KOZKREOLMRU systems out-performed CREOLEM2M for `mfe↔eng` MT, but both under-performed `kreyol-mt-pubtrain`, with the out-of-English direction performingly comparitively worse. (This is consistent with Rajcoomar's (2025) own finding that monolingual pre-training and two steps of fine-tuning has different effectiveness

direction	Team	Model	chrF++	BLEU
		creole → XX		
aoa→eng	**EHOW**	contrastive1	34.9	30.0
	EHOW	primary	19.3	17.6
	baseline	`kreyol-mt`	10.5	4.4
cri→eng	*baseline*	`kreyol-mt`	82.8	79.9
	EHOW	primary	82.0	78.2
fab→eng	**EHOW**	contrastive2	28.2	15.4
	EHOW	contrastive1	27.7	14.7
	EHOW	primary	12.0	1.2
	baseline	`kreyol-mt`	10.0	0.3
kea→eng	*baseline*	`kreyol-mt`	93.7	90.1
	EHOW	primary	93.4	90.0
mfe→eng	**KOZKREOL**	primary	46.7	25.6
	baseline	`kreyol-mt-pubtrain`	46.3	25.0
	baseline	CreoleM2M	33.5	12.8
pap→eng	**EHOW**	contrastive1	84.0	74.8
	EHOW	primary	76.3	64.8
	baseline	`kreyol-mt`	74.6	62.1
	baseline	CreoleM2M	56.7	37.1
pov→eng	*baseline*	`kreyol-mt`	87.7	82.8
	EHOW	contrastive1	81.0	74.0
	EHOW	primary	81.0	74.0
pre→eng	**EHOW**	contrastive2	56.6	40.3
	EHOW	contrastive1	55.3	40.5
	EHOW	primary	24.1	9.3
	baseline	`kreyol-mt`	9.9	0.3
		XX → creole		
eng→aoa	**EHOW**	contrastive2	33.2	24.4
	EHOW	contrastive1	33.0	23.5
	EHOW	primary	27.8	21.4
	baseline	`kreyol-mt`	8.7	12.4
eng→cri	*baseline*	`kreyol-mt`	80.2	76.5
	EHOW	contrastive1	78.1	73.6
	EHOW	primary	25.4	7.3
eng→fab	**EHOW**	contrastive2	26.0	5.1
	EHOW	contrastive1	25.5	7.7
	EHOW	primary	16.1	2.6
	baseline	`kreyol-mt`	6.6	0.8
eng→kea	*baseline*	`kreyol-mt`	91.4	87.5
	EHOW	contrastive1	90.1	85.5
	EHOW	primary	41.5	17.9
eng→mfe	*baseline*	`kreyol-mt-pubtrain`	49.7	28.7
	KOZKREOL	primary	43.1	18.6
	baseline	CreoleM2M	32.7	10.0
eng→pap	**EHOW**	contrastive1	76.7	62.2
	EHOW	primary	72.1	53.0
	baseline	`kreyol-mt`	65.6	48.8
	baseline	CreoleM2M	51.4	29.5
eng→pov	*baseline*	`kreyol-mt`	92.0	89.9
	EHOW	contrastive1	74.1	67.6
	EHOW	primary	31.7	12.3
eng→pre	**EHOW**	contrastive2	44.6	21.8
	EHOW	contrastive1	42.4	22.8
	EHOW	primary	26.4	5.4
	baseline	`kreyol-mt`	9.1	1.2

Table 2: Results of the **EHOW** and **KOZKREOLMRU** (abbreviated **KOZKREOL**) submissions for the unconstrained track. Systems are ordered by chrF++ score.

[12] Note that CREOLEM2M supports only two of the submitted language pairs: `mfe↔eng` and `pap↔eng`.

	mfe→eng		eng→mfe	
	FLORES	LALIT	FLORES	LALIT
baseline	57.3	50.8	49.1	46.2
primary	**67.7**	**70.2**	**57.7**	**68.9**

Table 3: Comparison of KOZKREOLMRU chrF scores with `kreyol-mt` on FLORES and LALIT test sets. Here *primary* is understood to refer to the version of the *primary* submitted systems that was not trained on FLORES `devtest` data, in the columns for FLORES. These are averaged sentence chrF scores across each set, consistent with Rajcoomar (2025). High scores are **bold**.

depending on language direction.) But despite its under-performance on the official evaluation, the KOZKREOLMRU systems significantly outperform `kreyol-mt-pubtrain` on the KOZKREOLMRU submitted test sets (by 8.0 chrF minimum). See Table 3, which compares chrF (Popović, 2015) scores of Rajcoomar's (2025) models with `kreyol-mt-pubtrain` performance on both FLORES and LALIT test sets.

Scores for **EHOW** are also in Table 2. As noted in Section 3.1, the EHOW team used the `kreyol-mt` model (trained on both public and private KREYÒL-MT data). For a fair comparison, we use this model as their baseline. However, since `kreyol-mt` was trained on portions of the public test sets for `cri`, `kea`, `pap`, `pre`, and `pov`, this results in inflation of scores for both EHOW systems and their baselines.

Recall from Section 4.2.1 that the EHOW *primary* and *contrastive1* submissions consisted mostly of merged models (the best model for each overall direction, and the best for each language pair, respectively). In cases where LLM post-editing improved on the *contrastive1* result, this was submitted as *contrastive2*. Scores tend to improve from *primary*→*contrastive1*→*contrastive2* for most language pairs. EHOW systems outperform the baseline in both translation directions for `pap`, `pre`, `fab`, and `aoa`. The baseline scores higher on `cri`, `kea`, and `pov`. However, note that, as mentioned in the previous paragraph, these results and the others for `cri`, `kea`, `pap`, `pre`, and `pov` may be obfuscated by dataset contamination, making it difficult to draw clear conclusions from EHOW scores. To address this issue, we set up a second round of evaluations for a subset of the EHOW submissions.

5.2.1 Reevaluation of EHOW systems

Table 4 contains the results for our accurate reevaluation of EHOW *primary* and *contrastive1* systems, avoiding data contamination. Avoiding this contamination was a challenge. The team fine-tuned some models from a `kreyol-mt` initialization, which was trained on some segments in the test set used to evaluate `kreyol-mt-pubtrain`. And for fine-tuning they used the set used to train `kreyol-mt-pubtrain`, which contains some overlap with the test set used to evaluate `kreyol-mt`. Noting that each KREYÒL-MT model's corresponding train and test sets had no overlap with each other, we decided to use the intersection of both `kreyol-mt` and `kreyol-mt-pubtrain` test sets to ensure we would not evaluate on any segments used in training. However, this intersection was incredibly small for some language pairs. Hence, we augmented any resulting test sets with fewer than 20 aligned sentences, by adding sentences that had originally been filtered out of the test sets during Robinson et al.'s (2024) test set cleaning processes. These extra segments were removed due to length or noise, so we cleaned them manually. In these decontaminated, augmented test sets, the smallest set was for `pov-eng` (with 23 aligned sentences), just as in the original test set (in which the set for this same language pair had 33 aligned sentences).

When we evaluate on decontaminated test sets, EHOW systems outperform the baseline in chrF++ for every language pair, except those involving `pap` and `pov`. See Table 4. (Note that we can conclude from Table 2 already that EHOW outperformed the baseline on `aoa` and `fab` directions, not shown in Table 4.)

It is worth mentioning here that the EHOW team found superior performance of their systems over the baseline for directions involving `pap` and `pov` when they used their own test sets—which better match the distribution of the additional training data the team curated and used for fine-tuning. See Table 5 for chrF (Popović, 2015) scores. The EHOW in-house test sets are not completely free from contamination, since they contain some of the synthetic data segments tha the team produced using the `kreyol-mt` model itself. However, this would ostensibly give the baseline mode an advantage, rather than the competitor models; and the EHOW test sets for `pap-eng` and `pov-eng` only contain 13% and 15% synthetic segments, respec-

direction	Model	chrF++	BLEU
creole → eng			
cri→eng	primary	39.2	22.0
	kreyol-mt	37.2	22.6
kea→eng	primary	61.0	43.9
	kreyol-mt	56.5	36.1
pap→eng	kreyol-mt	67.8	54.2
	primary	66.6	52.6
	contrastive1	63.1	47.2
pov→eng	primary	51.0	41.8
	kreyol-mt	43.0	27.4
	contrastive1	40.4	19.4
pre→eng	contrastive1	59.7	54.5
	primary	26.3	9.9
	kreyol-mt	5.8	0.1
eng → creole			
eng→cri	primary	35.7	24.3
	kreyol-mt	28.6	16.5
	contrastive1	27.6	12.2
eng→kea	contrastive1	50.0	27.8
	kreyol-mt	50.0	26.7
	primary	43.6	22.3
eng→pap	kreyol-mt	59.3	41.1
	contrastive1	58.4	41.4
	primary	46.8	27.2
eng→pov	kreyol-mt	29.3	8.3
	contrastive1	27.5	7.5
	primary	25.9	3.4
eng→pre	contrastive1	31.2	14.0
	primary	25.9	10.6
	kreyol-mt	9.0	1.3

Table 4: Results from reevaluating EHOW *primary* and *contrastive1* submissions on a decontaminated test set. Systems are ordered by chrF++ score.

tively. Hence we infer that the effects of this data contamination are minor, and can conclude with reasonable confidence that EHOW's own models for these language pairs would outperform the baseline model in the genres represented in the extra data they used.

Given all of this, we observe that that both unconstrained submissions (EHOW and KOZKREOLMRU) show a common pattern: employing new training datasets and different pre-trained models can expand Creole MT performance to new genres, even in cases when it does not significantly improve performance on pre-existing test sets or distributions.

6 Conclusions

The first shared task for Creole language MT convened submissions for a variety of Creole languages situated across the Caribbean, South America, and Africa. This convergence was made possible by an important acknowledgment: despite their differences, Creole languages may benefit from a united approach in developing new language technology solutions. Indeed, Creole languages are the fruit of similar sociohistorical developments, leading to shared linguistic patterns, and also have in common a paucity of corpora and MT systems. We received and evaluated four submissions for new MT systems, and two dataset submissions, representing 12 Creole languages total.

	XX→eng		eng→XX	
	pap	pov	pap	pov
baseline	39.5	29.8	38.8	20.1
primary	45.8	28.6	26.9	**44.2**
contrastive1	**67.6**	**46.2**	**49.5**	18.4

Table 5: Comparison of EHOW submitted models to the kreyol-mt baseline chrF on private EHOW test sets. Highest scores are **bold**.

We note several noteworthy observations from the contributions of the participants:

- **Creative data curation**: We observed a wide variety of approaches, including human translation, data augmentation, data up-sampling, back-translation, and synthetic data generation.
- **Harnessing linguistic information**: Submissions demonstrated the utility of linguistic considerations and relationships between languages (both between Creole languages with shared history and with relative languages).
- **Data-conscious methods**: Similarly, participants leveraged an assortment of algorithmic approaches to overcome data scarcity, including LoRa PEFT, partial freezing of layers, and model merging.
- **Adressal of directional challenges**: Participants noted that translation into a high-resource language tended to yield better results than translation into a Creole language.

Our evaluation led to some central lessons and takeaways. It was found that constrained systems, which attempted to boost MT performance for a particular language by post-training the kreyol-mt-pubtrain model on a subset of its own training data (pertaining to the language in question), did not result in significant performance improvements, even when authors searched across other training tactics and hyperparameters to maxi-

mize performance. This suggests that developing new datasets and using pre-trained models may be a more promising direction. Accordingly, our participants in the unconstrained systems track showed that such methods are often effective at improving results for some language pairs, and that even in cases where performance on the original test domain does not improve, new datasets and models can bring about expansion to new testing genres.

Takeaways and Future Ambitions The variety of training approaches developed and evaluated for this shared task provides valuable insights into the training of Creole MT systems. New data collected during the campaign can also be incorporated into future Creole MT datasets and models. In particular, this will allow us to train a new baseline model for the next iteration of the shared task. Re-training the baseline model will also give us the chance to ensure that all updated versions of KREYÒL-MT models online can be evaluated with the same test sets without contamination (a critical point for interpreting results).

We are proud that a number of this year's shared task organizers represent various Creole language-speaking communities. Our committee includes two L1 speakers of Jamaican Patois, one L1 speaker of Martinican Creole, one L2 speaker of Louisiana Creole, and one L2 speaker of Haitian. In the future, we hope to incorporate members from a broader diversity of Creole language communities, so that our efforts better serve the realistic needs of these communities.

Limitations

As noted throughout this paper, our primary limitation stemmed from issues with data contamination between the train and test sets for `kreyol-mt` and `kreyol-mt-pubtrain models`. This was an organizational failure on our part that will be rectified in future iterations of this work. A limitation not so easily overcome is simply that of genre homogeneity in the datasets for low-resource languages. As demonstrated by both unconstrained track participants, it was much easier to out-perform pre-trained baseline models on novel test sets than on existing test sets, likely due to correlations of genre and topic between training data and testing data. Though this is a significant limitation, it is one of the very problems that this shared task is intended to rectify. (Due to this year's progress, we now have more diverse datasets for Mauritian Creole, Belizean Kriol, and French Guianese Creole.)

Ethical Statement

Given the historical and ongoing marginalization of many Creole languages and their population of speakers (DeGraff, 2003), we stress that community engagement is crucial. To ensure resulting research in machine translation is in accordance with community wants and needs (Lent et al., 2022), with the goal of preserving community autonomy (Bird, 2020). That said, one common limitation of working in the low-resource space is over-reliance on religious domain data; we acknowledge the presence of data which may not be culturally relevant to Creole language speakers (Hershcovich et al., 2022; Mager et al., 2023).

Acknowledgments

This work received funding from the Inria "Défi"-type project COLaF. We would like to thank the WMT organizers for making a place for Creole languages and for the opportunity to host a shared task. We also thank all of the teams who submitted systems and datasets.

References

Ananya Ayasi. 2025. Krey-All WMT 2025 CreoleMT System Description: Language Agnostic Strategies for Low-Resource Translation. In *Proceedings of the Tenth Conference on Machine Translation*.

Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. 2011. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, 26(1):5–42.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Christopher Clarke, Roland Daynauth, Jason Mars, Charlene Wilkinson, and Hubert Devonish. 2024. GuyLingo: The Republic of Guyana Creole Corpora. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 792–798, Mexico

City, Mexico. Association for Computational Linguistics.

Raj Dabre and Aneerav Sukhoo. 2022. Kreol-MorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Raj Dabre, Aneerav Sukhoo, and Pushpak Bhattacharyya. 2014. Anou Tradir: Experiences In Building Statistical Machine Translation Systems For Mauritian Languages – Creole, English, French. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88, Goa, India. NLP Association of India.

Michel DeGraff. 2003. Against creole exceptionalism. *Language*, 79(2):391–410.

Guillaume Fon Sing. 2017. Creoles are not typologically distinct from non-Creoles. *Language Ecology*, 1(1):44–74.

Charles Gilman. 1986. African Areal Characteristics: Sprachbund, not Substrate? *Journal of Pidgin and Creole Languages*, 1(1):33–50.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Roger M. Keesing. 1988. *Melanesian Pidgin and the Oceanic Substrate*. Stanford University Press.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1317–1327.

Silvia Kouwenberg and John Victor Singler. 2009. *The handbook of pidgin and creole studies*. John Wiley & Sons.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

William Lewis. 2010. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation*, St Raphael, France.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Ludovic Mompelat. 2025. Ludovic mompelat wmt 2025 creolemt systems description : Martinican creole and french. In *Proceedings of the Tenth Conference on Machine Translation*.

Salikoko S. Mufwene. 2001. Creolization is a social, not a structural, process. In Ingrid Neumann-Holzschuh and Edgar W. Schneider, editors, *Degrees of Restructuring in Creole Languages*, Creole Language Library, page 65. John Benjamins Publishing Company.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *ArXiv*, abs/1912.03444.

Robert Antoine Papen. 1978. *The French-based Creoles of the Indian Ocean: an Analysis and Comparison.* University of California, San Diego.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Hemkeshsing Y. Rajcoomar. 2025. KozKreolMRU WMT 2025 CreoleMT System Description: Koz Kreol: Multi-Stage Training English-Mauritian Creole MT. In *Proceedings of the Tenth Conference on Machine Translation*.

Nathaniel Robinson. 2025. JHU WMT 2025 CreoleMT System Description: Data for Belizean Kriol and French Guianese Creole MT. In *Proceedings of the Tenth Conference on Machine Translation*.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Jacqueline Rowe, Ona de Gibert, Mateusz Klimaszewski, Coleman Haley, Alexandra Birch, and Yves Scherrer. 2025. EdinHelsOW WMT 2025 CreoleMT System Description: Improving Lusophone Creole Translation through Data Augmentation. In *Proceedings of the Tenth Conference on Machine Translation*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. *arXiv preprint arXiv:2010.06354*.

Aileen Joan Vicente, Theresse Faith Amamampang, Dunn Dexter Lahaylahay, and Charibeth Cheng. 2024. ChavacanoMT: A Corpus and Evaluation of Neural Machine Translation for Philippine Creole Spanish.

Findings of WMT 2025 shared task on Low-resource Indic Languages Translation

Partha Pakray NIT Silchar
Reddi Mohana Krishna NIT Silchar
Santanu Pal Wipro AI Lab45
Advaitha Vetagiri NIT Silchar
Sandeep Kumar Dash NIT Mizoram
Arnab Kumar Maji North-Eastern Hill University
Saralin A. Lyngdoh North-Eastern Hill University
Lenin Laitonjam NIT Mizoram
Anupam Jamatia NIT Agartala
Koj Sambyo NIT Arunachal Pradesh
Ajit Das Bodoland University
Riyanka Manna Amrita Vishwa Vidyapeetham Amaravati

Abstract

This study proposes the results of the low-resource Indic language translation task organized in collaboration with the Tenth Conference on Machine Translation (WMT) 2025. In this workshop, participants were required to build and develop machine translation models for the seven language pairs, which were categorized into two categories. Category 1 is moderate training data available in languages i.e English–Assamese, English–Mizo, English-Khasi, English–Manipuri and English–Nyishi. Category 2 has very limited training data available in languages, i.e English–Bodo and English–Kokborok. This task leverages the enriched IndicNE-corp1.0 dataset, which consists of an extensive collection of parallel and monilingual corpora for north eastern Indic languages. The participant results were evaluated using automatic machine translation metrics, including BLEU, TER, ROUGE-L, ChrF, and METEOR. Along with those metrics, this year's work also includes Cosine similarity for evaluation, which captures the semantic representation of the sentence to measure the performance and accuracy of the models. This work aims to promote innovation and advancements in low-resource Indic languages.

1 Introduction

The Indic MT Shared Task, first organized alongside the Eighth Conference on Machine Translation (WMT) 2023[1] (Pal et al., 2023), demonstrated the critical need for sustained research attention toward low-resourced languages. That inaugural effort not only revealed the untapped potential for advancing machine translation in these linguistic contexts but also provided a robust methodological and collaborative foundation for future work. Motivated by its impact, the task was expanded and refined in the Ninth Conference on Machine Translation (WMT) 2024[2] (Pakray et al., 2024), drawing broader participation and richer system diversity. Building upon these successive advancements, the 2025 edition[3] has emerged as the most successful iteration to date—surpassing previous years in both scale and the quality of contributions—further cementing the task's role as a driving force in low-resource MT research.

India's linguistic landscape is remarkably diverse, encompassing hundreds of languages spoken across its regions. While 22 languages are officially recognized under the Eighth Schedule of the Indian Constitution and receive governmental support in terms of infrastructure, research, and funding, many others—often spoken by indigenous and minority communities—remain excluded from such provisions. These low-resource languages frequently lack standardized orthographies, adequate lexical resources (e.g., corpora), and formal linguistic documentation. Limited institutional support, declining intergenerational transmission, and minimal access to modern technologies further threaten their preservation and vitality.

To address these challenges, our initiative is dedicated to revitalizing and documenting low-resource Indic languages through targeted, technology-driven solutions. Building upon the success of the Indic MT Shared Tasks at WMT 2023 and WMT 2024, which focused on four language pairs (En-

[1] https://www2.statmt.org/wmt23/indic-mt-task.html

[2] https://www2.statmt.org/wmt24/indic-mt-task.html

[3] https://www2.statmt.org/wmt25/indic-mt-task.html

Proceedings of the Tenth Conference on Machine Translation, pages 532–553
November 8-9, 2025

glish–Assamese, English–Mizo, English–Khasi, and English–Manipuri) using the enriched IndicNE-Corp1.0 dataset, we have expanded the scope in 2025 to seven language pairs. These are divided into two categories: (1) Languages with moderate amounts of training data: English–Assamese, English–Mizo, English–Khasi, English–Manipuri, and English–Nyishi. (2) Languages with very limited training data: English–Bodo and English–Kokborok.

This year's task, which has already surpassed prior editions in scale and participation, continues to drive innovation in machine translation and NLP, developing solutions specifically adapted to the unique linguistic and resource constraints of low-resource Indic languages. The Indic MT Shared Task initiative is also committed to safeguarding India's rich linguistic diversity and cultural heritage by strengthening the rights and identities of minority language communities. Leveraging state-of-the-art technologies, it seeks to advance the capabilities of low-resource Indic languages, enabling them not only to survive but to flourish within today's increasingly digital and interconnected landscape.

The task is evaluated using a wide range of metrics, integrating automatic lexical evaluation metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015). In addition to the standard evaluation metrics, we used cosine similarity using Sentence-BERT(all-mpnet-base-v2) (Reimers and Gurevych, 2019) to compute the semantic similarity between the predicted and reference sentences in the English language direction. By combining traditional and semantic evaluation methods, this approach provides a thorough assessment of the performance and accuracy of translation systems.

2 Languages

This section is divided into two parts based on the availability of training data: Category 1 includes languages with moderate resources, and Category 2 includes languages with very limited resources. In the WMT 2024 edition of the shared task, the focus was limited to four languages: Assamese, Mizo, Khasi, and Manipuri. Building upon that foundation, the WMT 2025 task expanded the linguistic coverage by including three additional languages: Nyishi, Bodo, and Kokborok. This expansion reflects our ongoing commitment to improving the representation of under-resourced languages in machine translation research and to broadening the scope of technological inclusion for more linguistically marginalised communities.

2.1 Category 1: (Moderate Training Data Available)

2.1.1 The Assamese Language

Assamese, a member of the Indo-Aryan(Wikipedia contributors, 2025a) branch of the Indo-European language family, is primarily spoken in the northeastern Indian state of Assam. It holds official status in the state and functions as a vital lingua franca, bridging communication across the region's diverse ethnic communities. Recognized as one of India's 22 scheduled languages, Assamese occupies a prominent position within the nation's multilingual framework.

With literary roots dating back to the early medieval era, Assamese(Mahanta, 2012) boasts a long-standing and vibrant cultural heritage. The script currently used for the language evolved from the ancient Brahmi script, reflecting centuries of historical development. However, in the contemporary digital age, Assamese confronts a number of challenges, particularly with regard to language technology. The creation and advancement of computational resources for Assamese are essential, not only to support its practical use in modern contexts but also to ensure its continued vitality and preservation in the face of rapid technological change.

2.1.2 The Mizo Language

Mizo, belonging to the Tibeto-Burman (Wikipedia contributors, 2025d) branch of the Sino-Tibetan language family, is primarily spoken in the northeastern Indian state of Mizoram. It serves as the main medium of communication among the Mizo community and is also used by various ethnic groups across neighbouring regions such as Manipur, Tripura, Assam, and even in parts of Myanmar and Bangladesh. Known for its tonal structure and distinct phonological characteristics, Mizo stands out as a linguistically unique language within the Tibeto-Burman group.

The Mizo language (Zothanliana, 2021) is deeply rooted in a vibrant oral tradition that encompasses folklore, songs, and storytelling forms of expression that preserve and reflect the community's cultural identity. The written form of the language began to take shape in the late 19th century, when Christian missionaries

introduced the Roman script, enabling systematic transcription and laying the groundwork for written literature. Since then, Mizo has developed a strong literary presence, with works spanning traditional poetry to contemporary prose. Nonetheless, like many regional languages, Mizo faces significant challenges in the digital era, especially concerning its representation in language technology and digital communication platforms. Addressing these issues is crucial for both the preservation and advancement of the language.

2.1.3 The Khasi Language

Khasi, a language of the Austroasiatic family(Wikipedia contributors, 2025c), is predominantly spoken in the central and eastern regions (Rynjah and Lyngdoh, 2023) of Meghalaya. Prior to 1813, Khasi lacked a writing system. Between 1813 and 1814, the Bengali script was adopted for translating the Bible into Khasi, a decision influenced by the relatively high literacy in Bengali during that period. By 1816, translated excerpts from the Gospel of Matthew had been printed and circulated among Khasi speakers proficient in Bengali. However, a significant shift occurred in 1841 with the arrival of a Welsh missionary, who introduced the Roman script. This led to translations being made in the Sohra (Cherra) dialect, which later became the basis for standardized written Khasi.

Khasi is marked by notable dialectal diversity. Grierson(Grierson, 1903) identified four primary dialects: Standard Khasi, Pnar (or Synteng), Lyngngam, and War. Acharya (Acharya, 1971) reaffirmed this classification, adding that additional sub dialects such as Bhoi, spoken in Meghalaya's northern plains, also exist. Expanding on these observations, Bareh(Bareh, 1977) provides a detailed account of Khasi dialects, classified primarily by their geographical distribution:

- Amwi (southern Jaiñtia hills),
- Shella and Warding (southern Khasi hills),
- Myriaw, Nongkhlaw, Nongspung, Maram, and Mawiang (mid-western Khasi hills),
- Cherra (mid-southern Khasi hills),
- Mylliem, Laitlyngkot, Nongkrem, and Lyniong-Khasi (central Khasi hills),
- Jowai (central Jaintia hills),
- Bhoi (northeastern Khasi hills),
- Manar, Nongwah, and Jirang (northern Khasi hills),
- Khatarblang/Mawpran (mid-southern region),
- Nongstoin and Langrin (western Khasi hills).

Bareh also emphasizes that phonological variation exists within these groups. Among them, Amwi is considered a particularly distinct and conservative dialect. It is often viewed as more agglutinative and less intelligible to speakers of related dialects such as Jowai or Khatarblang. Despite its unique structure, Amwi speakers can generally understand and use neighbouring dialects for communication. Its linguistic features suggest strong retention of Mon-Khmer elements, especially in morphology and phonology.

Bareh (1977) organizes Khasi dialects into three broad branches:

1. **Eastern Dialects**:
   - Jowai (Central Highlands),
   - Amwi and the War dialects (southern region),
   - Bhoi Synteng (northern region).
2. **Central Dialects**:
   - Dialects such as Nongphlang, Cherra, Nongkrem, Mylliem, Nongspung, and others,
   - Northern varieties like Bhoi East (e.g., Mawrong, Bhoi Lymbong) and Bhoi West (e.g., Manar),
   - War Shala and Warding (southern region).
3. **Western Dialects**:
   - Nongstoin, Lyngam, and Langrin.

Within each branch, sub-dialects often display considerable variation, particularly in phonological patterns. Daladier(Daladier, 2002) notes that Khasi is a part of the Mon-Khmer subgroup of the Austroasiatic family retains conservative unwritten forms, especially in the War dialect areas. Pnar and War remain among the most prominent dialects, with War further subdivided into Nongtalang, Amvi, Tremblang, and Shella. The internal

classification of Pnar remains relatively unexplored, War dialects are further sub divided into War-Khasi and War-Jaintia, spoken in the southern regions of the Khasi and Jaiñtia hills.Grierson also provides foundational work on these dialects.

For the purposes of this shared task, we adopt the Sohra (Cherra) dialect as the standardized form of Khasi for translation. Its historical significance, combined with its widespread use in education, religion, and official discourse, makes it a practical and consistent choice. The formal adoption of the Roman script in 1841 further reinforced its position as the standard written variant, ensuring accessibility for both native speakers and learners across the Khasi-speaking population.

2.1.4 The Manipuri Language

Manipuri, also known as Meiteilon, is a Sino-Tibetan language primarily spoken in the northeastern Indian state of Manipur. It holds official status as one of the 22 scheduled languages of India and functions as a common medium of communication among diverse ethnic communities in the region, thereby playing a key role in facilitating both social interaction and cultural integration.

The language is marked by a rich literary tradition, with historical records indicating the use of written texts since ancient times. Manipuri employs both the indigenous Meitei Mayek script and, more recently, the Bengali script for written communication. Despite its cultural and historical significance, the language faces notable challenges related to preservation and modernization, particularly within the context of technological development and digital communication. The advancement of computational tools and linguistic resources is critical for ensuring the sustained vitality and broader accessibility of Manipuri in the digital age. In recent years, there has been an increasing academic and technological interest in developing natural language processing (NLP) tools tailored to low-resource languages, including Manipuri (Allen, 2003). However, several persistent issues continue to hinder progress (Gyanendro Singh et al., 2016). Chief among these is the scarcity of annotated corpora and high-quality linguistic datasets, which are crucial for training effective machine learning models. This lack of data significantly constrains the development of key NLP applications such as machine translation (Pal et al., 2023), sentiment analysis (Singh and Singh, 2017), and automatic speech recognition (Gyanendro Singh et al., 2016). Another major obstacle lies in the linguistic complexity of Manipuri itself. The language features a highly inflectional morphological system, posing difficulties for standard NLP models, which are typically optimized for morphologically simpler and better-resourced languages like English. Additionally, issues of script representation and lack of digital standardization complicate text processing, as existing tools often struggle with script conversion, normalization, and consistency across platforms.

Ongoing research is working to mitigate these challenges by building foundational linguistic resources, designing language-specific processing algorithms, and modifying existing NLP architectures to better accommodate the structural characteristics of Manipuri. Despite these promising developments, there remains a substantial gap between Manipuri and more digitally privileged languages a gap that must be addressed through sustained linguistic, technological, and policy-driven efforts.

2.1.5 The Nyishi Language

Nyishi(Kakum et al., 2023; Wikipedia contributors, 2025e), also known as Nyising, Leil, Aya, Nisi, Bangni-Bangru, or Akang, is a Tani language within the Sino-Tibetan family, spoken across eight districts in Arunachal Pradesh. It exhibits distinct typological features, including tonal phonology and context-dependent semantics.

The language uses a modified Roman script comprising seven vowels, eighteen consonants, consonant clusters, a glottal component, and a semivowel. Nyishi is tonal, employing rising, falling, and level tones to distinguish meaning an essential feature in its monosyllabic lexical system.

Syntactically, Nyishi follows a default Subject–Object–Verb (SOV) word order, with occasional Subject–Verb–Object (SVO) patterns. It blends isolating and agglutinative morphological traits. Words often carry multiple meanings based on context for instance, *taxy* may mean “squirrel”, “ginger”, or “animal lice”.

Gender is marked via suffixes rather than noun inflection, using forms like *kibu* (male dog) and *kine* (female dog). There are no plural markers or verb agreement for person or number. Negation is expressed uniformly through the particle *ma*.

Despite its cultural and linguistic value, Nyishi remains underrepresented in technological and computational domains. Expanding NLP efforts

and developing language resources are crucial for its digital preservation and broader accessibility.

2.2 Category 2: (Very Limited Training Data)

2.2.1 The Bodo Language

Bodo, also referred to as Boro (Wikipedia contributors, 2025b), is a member of the Bodo-Garo subgroup within the Tibeto-Burman branch of the Sino-Tibetan language family. It is primarily spoken by the Bodo people in the Bodoland Territorial Region (BTR) of Assam, India, with additional speaker communities in neighboring states such as Arunachal Pradesh, Meghalaya, and Nagaland, as well as in parts of Nepal and Bangladesh. According to the 2011 Census of India, the language is spoken by approximately 1.4 million people.

Bodo holds official recognition as one of the 22 scheduled languages of India(Bhattacharya, 1977), having been included in the Eighth Schedule of the Constitution in 2003. It also enjoys the status of an associate official language in Assam and is used as a medium of instruction in educational institutions within the BTR. The language was historically written in the Latin and Assamese scripts, but since 1975, the Devanagari script has been officially adopted.

Linguistically, Bodo is characterized as a tonal and agglutinative language. Its tonal system assigns semantic distinctions based on pitch, while its morphology supports the formation of complex words through the use of multiple affixes. The syntactic structure typically follows a Subject-Object-Verb (SOV) (Pathak et al., 2025) order, aligning with patterns common to Tibeto-Burman languages.

The Bodo literary tradition has developed (Boro, 2021)significantly in the modern period, particularly following the establishment of the Bodo Sahitya Sabha in 1952, which has been central to efforts in language standardization, publication, and literary development. Today, Bodo literature encompasses diverse genres, including poetry, fiction, and drama, reflecting the sociocultural life of its speakers.

Despite its official status and growing corpus, Bodo remains relatively under-resourced in the digital and computational linguistic domains. Continued initiatives in documentation, corpus building, and NLP development are essential to ensure its sustained vitality and technological integration.

2.2.2 The Kokborok Language

Kokborok, also known as Tripuri, is a Tibeto-Burman language belonging to the Bodo-Garo subgroup, primarily spoken in the Indian state of Tripura and parts of southern Assam, Mizoram, and the Chittagong Hill Tracts of Bangladesh (Debbarma et al., 2012; Nagaraja, 2015). It serves as a lingua franca among indigenous communities, including the Tripuri, Reang, Jamatia, Debbarma and other Borok tribes. According to the 2011 Census of India (Census of India, 2011), Kokborok has approximately 800,000 speakers in India, with additional speakers in Bangladesh, though exact figures for the latter are less documented. Though Kokborok is one of the official language of Tripura, but in urban areas like Agartala, there is a noticeable shift toward Bengali due to its dominance in administration, education, and media.

Linguistically, Kokborok has a rich phonological system, featuring six monophthong vowels and a consonant inventory that includes stops, nasals, fricatives, and approximants. Historically tonal, with pitch distinctions marking lexical meaning, the language is undergoing phonological simplification, particularly in urban settings, due to prolonged contact with non-tonal languages like Bengali and Hindi. Morphologically, Kokborok is agglutinative, employing affixation and compounding to form complex words(Hoque, 2014). Its syntax follows a Subject-Object-Verb (SOV) order, with a robust case system and verbal inflections for tense, aspect, mood, and person, reflecting its Tibeto-Burman roots.

Kokborok is written in both Roman and Bengali scripts, creating challenges for standardization and literacy efforts. Since 1979, Kokborok has been recognized as an official language in Tripura and is integrated into primary and secondary education curricula, with efforts to develop textbooks and teaching materials (Roy et al., 2022). The language remains a cornerstone of Borok cultural identity, expressed through oral traditions, folklore, songs, and ritual practices, such as those tied to festivals like Garia, Kharchi, Ker, and Hojagiri(Jacquesson, 2003).

Despite its cultural significance, UNESCO's Atlas of the World's Languages in danger classifies Kokborok as "vulnerable", reflecting threats from language shift and limited intergenerational transmission in urban areas(UNESCO, 2010). Revitalization efforts are ongoing, including curricu-

lum development, cultural festivals, and media programming like radio broadcasts and local television. Computational innovations, such as Linear Predictive Cepstral Coefficients (LPCC) for vowel recognition, support language documentation and preservation(Debbarma, 2012). Advocacy for Kokborok's inclusion in the Eighth Schedule of the Indian Constitution continues, emphasizing its linguistic and political significance for the Borok people.

3 Low-Resource Indic Language Translation 2025 Shared Task

3.1 Overview and Task Description

Following the success of the "Shared Task: Low-Resource Indic Language Translation" at WMT 2024, which attracted widespread international participation, the initiative will continue as part of the Tenth Conference on Machine Translation (WMT 2025). While recent advancements in machine translation (MT), particularly through multilingual modelling and transfer learning, have led to significant performance gains, developing effective MT systems for low-resource languages remains a critical challenge. This difficulty primarily stems from the limited availability of high-quality parallel corpora, which are essential for training robust and accurate translation models. The shared task aims to address this gap by fostering research and collaboration in low-resource Indic MT and promoting the creation and evaluation of translation systems for linguistically diverse and underrepresented languages.

The WMT 2025 Indic Machine Translation Shared Task aims to tackle the persistent challenges of low-resource translation by focusing on a diverse set of Indic languages drawn from multiple language families. This year, the task is organized around two categories based on the availability of training data.

- **Category 1** includes language pairs with moderate amounts of parallel data: English ⇔ Assamese, English ⇔ Mizo, English ⇔ Khasi, English ⇔ Manipuri, and English ⇔ Nyishi.
- **Category 2** consists of language pairs with extremely limited training resources: English ⇔ Bodo and English ⇔ Kokborok. By highlighting both moderately and severely resource-constrained languages, the task encourages the development of adaptable and data-efficient machine translation approaches capable of addressing the varying degrees of resource scarcity.

3.2 Categories

This year's task features two main categories based on the availability of training data:

3.2.1 Category 1: Moderate Training Data

- English ⇔ Assamese (en⇔as)
- English ⇔ Mizo (en⇔lus)
- English ⇔ Khasi (en⇔kha)
- English ⇔ Manipuri (en⇔mni)
- English ⇔ Nyishi (en⇔njz)

3.2.2 Category 2: Very Limited Training Data

- English ⇔ Kokborok (en⇔trp)
- English ⇔ Bodo (en⇔bodo)

3.3 Goal

This shared task goal is to build machine translation systems that can generate accurate translations regardless of the limitations of limited data availability. Participants are motivated to explore different innovative approaches, including:

- **Leveraging Monolingual Resources**: Utilizing monolingual corpora to improve the performance of translation systems.
- **Multilingual Strategies**: Exploring cross-lingual transfer techniques to support translation for under-resourced language pairs.
- **Cross-lingual Transfer Learning**: Employing models pretrained on high-resource languages and adapting them to low-resource scenarios.
- **Novel Methodologies**: Applying cutting-edge or customized approaches designed specifically for data-scarce environments.

3.4 Data

3.4.1 Training

This WMT 2025 Indic Machine Translation Shared Task leverages a dataset that consists of both parallel and monolingual corpora for Assamese, Khasi, Mizo, Manipuri, Nyishi, Bodo and Kokborok taken from the IndicNE-corp1.0 dataset.

3.4.2 Testing

For the testing section, we have created 2000 language pair sentences for each of the following language pairs:

- English ⇔ Assamese (en⇔as)
- English ⇔ Mizo (en⇔lus)
- English ⇔ Khasi (en⇔kha)
- English ⇔ Manipuri (en⇔mni)
- English ⇔ Nyishi (en⇔njz)
- English ⇔ Kokborok (en⇔trp)
- English ⇔ Bodo (en⇔bodo)

Out of these 2000 sentences, the first 1000 are provided in English, the participant needs to translate them into the specific target(Indic) language, and the remaining 1000 are given in the target language and are to be translated to English.

3.5 Evaluation

All the machine translation systems that are submitted by the participants are evaluated using automatic assessments to ensure balanced analysis of the translation systems. Automatic evaluation is being carried out by the following metrics such as BLEU, TER, ROUGE-L, ChrF and METEOR. Along with those metrics, this year's work also includes Cosine similarity using sentence transformer (all-mpnet-base) model based embeddings for evaluation, which captures the semantic representation of the sentences in the English language direction to measure the performance and accuracy of the models.

4 Dataset

4.1 Training

The dataset for the WMT 2024 Shared Task on Low-Resource Indic Language Translation is primarily based on the IndicNE-Corp1.0 dataset[4]. This corpus was built by aggregating datasets from previous research, including significant contributions from (Laskar et al., 2020) (Laskar et al., 2022), (Khenglawt et al., 2022), and (Laitonjam and Ranbir Singh, 2021). The compiled datasets encompass both parallel and monolingual corpora across four languages: Assamese, Mizo, Khasi, and Manipuri.

In earlier studies, we focused on developing parallel and monolingual corpora for English ⇔ Assamese (en⇔as) (Laskar et al., 2020, 2022), English ⇔ Mizo (en⇔lus) (Khenglawt et al., 2022), English ⇔ Khasi (en⇔kha) (Laskar et al., 2021), and English ⇔ Manipuri (en⇔mni) (Laitonjam and Ranbir Singh, 2021). The data was sourced from a variety of online platforms, including the Bible, multilingual dictionaries (such as Xobdo and Glosbe), multilingual question papers, PMIndia (Haddow and Kirefu, 2020), web pages, blogs, and online newspapers.

Table 1 shows the detailed statistics of the parallel datasets used for training and validation for each language pair.

Type	Sentences	Tokens (eng)	Tokens (target)
Assamese	54,000	1,033,580	878,466
Mizo	50,000	981,513	1,044,077
Khasi	26,000	778,689	948,853
Manipuri	23,687	422,522	357,524
Nyishi	60,000	337,887	323,876
Bodo	15,215	228,219	204,926
Kokborok	2,269	55,634	51,268

Table 1: Parallel data statistics for train and validation.

4.2 Testing

The testing dataset for the 2024 shared task was meticulously curated to present a substantial challenge beyond previous years' datasets. It comprised 1000 samples for each language pair, spanning four distinct and diverse domains: News, Travel, Sports, Entertainment, and Business. This domain-specific distribution aimed to comprehensively evaluate models' performance across varied and complex linguistic contexts, reflecting real-world translation demands. A collaborative approach was employed to create these testing samples, involving four specialized teams, each dedicated to one domain. These teams were provided 1000 English sentences, which they translated into their assigned target languages. The translation teams were instructed to maintain high fidelity to the source material while ensuring the translations were idiomatic and contextually appropriate for each domain.

The test set release process was intentionally staged to introduce additional complexity and rigour. In the first phase, 500 English sentences were released, requiring participants to translate these into the target languages. This forward trans-

[4] https://data.statmt.org/wmt23/indic-mt/

Language Pair	Entertainment	Sports	Healthcare	Travel	Political
English → Assamese	400	400	400	400	400
English → Mizo	400	400	400	400	400
English → Khasi	400	400	400	400	400
English → Manipuri	400	400	400	400	400
English → Nyishi	400	400	400	400	400
English → Bodo	400	400	400	400	400
English → Kokborok	400	400	400	400	400

Table 2: Domain-wise distribution of the 2025 test dataset across all language pairs. Each pair contains 2000 sentences, distributed evenly over five domains.

lation task required participants to demonstrate their models' proficiency in capturing nuances and domain-specific terminology in the target languages. In the second phase, 500 sentences in the target languages were provided, requiring translation back into English. This reverse translation task assessed the models' ability to accurately render the meaning, tone, and subtleties of the original sentences in English, thus testing bidirectional translation capability. The combined forward and reverse tasks aimed to evaluate the accuracy, fluency, and idiomatic correctness of the translations. The careful selection of diverse domains and the structured release of the test set were intended to challenge the generalization capabilities of the participating models. The goal was to ensure that only the most robust models, capable of handling a wide range of real-world scenarios, would excel.

This approach ensures a rigorous and multifaceted evaluation, capturing the subtleties of each language pair's translation performance across different domains.

5 Participants and System Descriptions

Language Pair	Submissions
English - Assamese	17 (primary), 18 (contrastive)
English - Mizo	5 (primary), 9 (contrastive)
English - Khasi	6 (primary), 19 (contrastive)
English - Manipuri	11 (primary), 7 (contrastive)
English - Nyishi	6 (primary), 12 (contrastive)
English - Bodo	6 (primary), 7 (contrastive)
English - Kokborok	3 (primary), 7 (contrastive)

Table 3: Number of participants in the low-resource Indic language translations

In this WMT 2025 Indic MT Shared Task, a total of 17 teams, as illustrated in the Table 4, registered and contributed for this year which is a huge improvement over the last year. We gathered the outputs produced by participant systems, including both primary and contrastive submissions.

A3-108: This team (Yadav and Shrivastava, 2025) system focused on translation for low-resource language pairs, combining a phrase-based SMT framework with subword segmentation through multiple BPE merge operations (500–3000 merges). Their approach involved concatenating and deduplicating segmented bitext to enhance vocabulary coverage and reduce sparsity, supported by KenLM-trained target-side language models. They submitted results for four English–X pairs: Nyishi, Manipuri, Khasi, and Assamese.

AkibaNLP-TUT: The AkibaNLP-TUT (Hamada et al., 2025) team tackled the WMT25 IndicMT task with Transformer-based models implemented in Fairseq. Their approach combined official parallel datasets with additional Bengali–English and Assamese monolingual corpora. A key technique was language-specific word-level noise injection to enhance robustness in low-resource settings, complemented by back-translation to augment English→X training data.

ANVITA: This team (Sivabhavani et al., 2025)submitted systems for three low-resource Indic languages Nyishi, Khasi, and Kokborok covering both primary and contrastive tracks. Their models were built using transfer learning, fine-tuning public pre-trained architectures such as byt5-base and nllb-200-distilled-600M with selective vocabulary expansion and targeted post-editing. The primary submissions relied on organizer-provided datasets, while the contrastive runs incorporated data augmentation through back-translation, sentence concatenation, and proprietary crawled resources. Language-specific strategies included leveraging Bodo data for Kokborok and tailoring vocabulary for Khasi.

BibaoMT: This team submission explored a

Team Name	Name of University/Lab/Industry/Group
A3-108	IIIT Hyderabad
AkibaNLP-TUT	Toyohashi University of Technology / NLP Lab.
ANVITA	CAIR
BilbaoMT	University of the Basque Country
BVSLP	Banasthali Vidyapith
CITK_MT	Central Institute of Technology Kokrajhar
DELAB-IIITM	Indian Institute of Information Technology, Senapati, Manipur
DoDS-IITPKD	Indian Institute of Technology, Palakkad, Kerala
DPKM	Dynamic Partial Knowledge Model Group
Hope for best	University of Delhi
JU-NLP	Jadavpur University
MT@HLT-BLR_Amrita	Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India
NLPTng-NITAP	National Institute of Technology Arunachal Pradesh
RBG-AI	Resilience Business Grids LLP
SRIB-NMT	Samsung Research Institute Bangalore
Trasnformers	Centre for Development of Advanced Computing (C-DAC), Pune
TranssionMT	Transsion Translation

Table 4: The following table provides an overview of the teams registered for the low-resource Indic language translation task at WMT 2025.

lightweight neural MT model with just 22.4M parameters (280 MB) to tackle low-resource English-to-Indic translation. Their approach used a two-stage training pipeline: multilingual pretraining on seven task languages plus three high-resource languages, followed by fine-tuning on target languages. Training data combined official shared-task corpora with NLLB-mined bitexts, Samanantar, HPLT Bengali–English and Hindi–English, and OpenSubtitles Spanish–English datasets, enabling efficient translation despite limited resources.

BVSLP: The BV-SLP(Joshi et al., 2025) team developed MT systems for five language pairs: English⇔Assamese, English⇔Manipuri, and English→Bodo. Their pipeline integrates a rule-based named entity recognition and translation module prior to NMT training, handling organisation and location names via translation or transliteration from a knowledge base. After preprocessing, byte pair encoding (BPE) was applied to prepare data for Transformer-based NMT training, enabling improved handling of named entities in low-resource scenarios.

CITK-MT: The CITK-MT(Wary et al., 2025) team proposed an end-to-end NMT system targeting English→Bodo translation, leveraging a Seq2Seq model with GRU-based encoder–decoder layers and Bahdanau attention to enhance contextual alignment. Their pipeline included extensive data preprocessing, careful hyperparameter tuning (embedding size, hidden units, dropout), and training on Google Colab with NVIDIA A100 GPUs. The system was optimized using early stopping and evaluated via BLEU scores, demonstrating a focused approach for low-resource language translation.

DELAB-IIITM: The DELAB-IIITM(Oinam and Saharia, 2025) team addressed low-resource Indic translation for English⇔Assamese and English⇔Manipuri by fine-tuning the NLLB-200 multilingual model with synthetic parallel data augmentation. They generated synthetic corpora by leveraging bilingual pairs (Manipuri–English, Assamese–English) to create additional data for target languages, yielding 77K sentences after strict data cleaning. Fine-tuning employed the Seq2SeqTrainer framework with the Adafactor optimizer (2e-5 learning rate) and two training epochs, alongside careful train–test splits to mitigate overfitting. Evaluation showed notable BLEU score gains over baseline NLLB models across most directions (e.g., mni-as 0.45 vs. 0.10 baseline).

DoDS-IITPKD: The DoDS IIT Palakkad (Khongthaw et al., 2025) team tackled low-resource Indic language translation by participating with four languages: Khasi, Mizo, Assamese (Category-1) and Bodo (Category-2). Their primary system fine-tuned facebook/nllb-200-distilled-600M for English⇔Khasi and English⇔Mizo, while IndicTrans2 was used for Assamese and Bodo. For the contrastive system, training data was expanded with external corpora

such as PMINDIA and Google SMOL, enabling broader coverage. Both systems applied Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning within the Hugging Face Transformers + PEFT framework, along with language-specific tagging for preprocessing. This modular design balanced translation quality with computational efficiency.

DPKM: The DPKM(Kumar et al., 2025) team presents a low-resource NMT approach for English–Kokborok and English–Bodo translation, leveraging the LLaMA2-8B model with LoRA-based parameter-efficient fine-tuning. Their pipeline involves pretraining on monolingual Kokborok and English corpora (70k Kokborok / 30k English for Kokborok, and 350k Kokborok / 125k English for Bodo) prior to instruction tuning using WMT25 datasets converted into Alpaca format to suit instruction-following objectives. The fine-tuning method integrates LoRA adapters to minimize training overhead on large models.

Hope for best: This team deployed pre-trained IndicTrans2 transformer models for English–Assamese translation without additional fine-tuning, prioritizing fast deployment under CPU-only constraints. They formatted inputs with language tags and applied minimal preprocessing, achieving balanced quality with beam search and batch inference. Their system highlights how off-the-shelf multilingual models can still perform competitively in low-resource shared tasks.

JU-NLP: The JUNLP(Acharya et al., 2025) team addressed English ⇔ Assamese, Mizo, Manipuri, and Bodo translation by fine-tuning multilingual NLLB and IndicTrans2 models using parameter-efficient methods like LoRA and DORA. Their pipeline featured rigorous preprocessing, including deduplication, script harmonization, and alignment filtering to improve data quality. Evaluation on WMT datasets showed competitive BLEU/ChrF scores despite low-resource constraints.

MTHLT-BLR_Amrita: This team(Sheshadri and Gupta, 2025) tackled Assamese and Bodo to English translation using IndicTrans2 enhanced with a novel Representation Fine-tuning (ReFT) method, inserting lightweight modules into encoder layers for targeted adaptation. They optimized ReFT hyperparameters via Bayesian search with Optuna and fine-tuned only 0.5M parameters to avoid overfitting. Experiments on WMT data demonstrated competitive BLEU scores under strict low-resource constraints.

NLPTng-NITAP: The team from NIT Arunachal Pradesh addressed English ⇔ Nyishi translation using the mBART50 multilingual model with Task-Adaptive Fine-Tuning (TAFT). They introduced a custom Nyishi token (<nyi_IN>) and performed full model fine-tuning on WMT25 parallel corpora, leveraging language prefixing for direction control. Their method demonstrates effective transfer learning for new Indic languages under severe low-resource constraints.

RBG-AI: The RBG-AI(H and Ptaszynski, 2025) team developed a multilingual translation pipeline using the MADLAD-400 T5-based model, optimized for both high- and low-resource languages. Their approach employed 4-bit quantization to reduce memory usage and speed up inference on RTX3090 hardware without compromising translation quality. The system incorporated language-specific tags and beam search decoding to improve fluency and directionality. This design balanced translation accuracy with deployment efficiency, suitable for edge or resource-limited environments.

SRIB-NMT: SRIB-NMT participated in the WMT-25 Low-Resource Indic MT challenge with contrastive submissions for four language pairs: English–Assamese, English–Mizo, English–Khasi, and English–Manipuri. Their system used pre-trained NLLB models combined with LoRA fine-tuning to efficiently adapt to low-resource settings. By leveraging cross-lingual transfer techniques, they achieved notable gains in SacreBLEU on blind test sets. The submission highlights parameter-efficient adaptation strategies for multilingual translation tasks.

Transformers: This team(Gupta et al., 2025) developed NMT models for English to Assamese, Bodo, and Manipuri using the OpenNMT framework with a Transformer-based encoder–decoder architecture. Their approach included extensive pre-processing tokenisation, BPE segmentation, and vocabulary generation along with transfer learning from Samanantar v2 to boost low-resource performance. The models were fine-tuned on WMT25 Indic datasets and evaluated with BLEU and perplexity metrics. Deployment was optimised through CTranslate2 for efficient runtime translation on GPUs.

TranssionMT: This team employed a dual-model strategy using IndicTrans2_1B and NLLB_3.3B for low-resource Indic translation. Their system applied cross-iterative back-translation of monolingual data to create high-quality pseudo-parallel corpora and semantic filtering (all-mpnet-base-v2) to enhance domain similarity. Rigorous data cleaning removed noise like URLs and untranslated segments, ensuring improved training quality. The final translations combined outputs from both models to achieve optimal results.

6 Results and Discussion

The results of the WMT 2025 Indic Machine Translation (MT) Shared Task are illustrated in the tables below. For clarity, results are reported separately for each language pair and direction: Assamese–English (as–en) in Table 5 and English–Assamese (en–as) in Table 6. Similarly, results for English–Manipuri (en–mni) and Manipuri–English (mni–en) are presented in Tables 7 and 8, respectively. The English–Khasi (en–kha) and Khasi–English (kha–en) directions are summarized in Tables 9 and 10. For English–Mizo (en–lus) and Mizo–English (lus–en), results are provided in Tables 11 and 12. The English–Nyishi (en–njz) and Nyishi–English (njz–en) results are shown in Tables 14 and 13. Similarly, results for and English–Bodo (en–bodo) and Bodo–English (bodo–en) are reported in Tables 15 and 16. Finally, results for English–Kokborok (en–trp) and Kokborok–English (trp–en) are detailed in Tables 17 and 18. Each table lists the participating systems in descending order of performance, along with their respective evaluation scores. This section presents the evaluation scores of the participants and their submitted system outputs and corresponding papers. Although participants submitted results for both primary and contrastive systems, only the primary system results are highlighted in the corresponding tables.

An evaluation of the quantitative results was performed using metrics like BLEU, METEOR, ROUGE-L, ChrF, and TER. BLEU measures the precision of n-grams in candidate translations relative to reference translations. TER quantifies the number of edits required to transform the candidate translation into the reference. ROUGE-L evaluates the longest common subsequence between the candidate and reference, emphasizing recall-oriented aspects of translation quality. ChrF computes the character n-gram F-score, providing sensitivity to morphological variations. METEOR combines precision, recall, and synonym matching to capture translation adequacy and fluency. In addition to the traditional statistical metrics this year's evaluation also incorporates semantic similarity based on cosine similarity between sentence embeddings generated by the all-mpnet-base model for the Indic language to English direction.

Discussion

For the Assamese language, the team TranssionMT achieved a higher BLEU score of 23.20 in primary mode and 22.41 in contrastive mode for the as-en direction with cosine similarity of 0.92 in primary mode of evaluation. For the en-as direction, this team also achieved a higher score in both primary and contrastive with BLEU scores of 20.97 and 67.50, respectively. This team employed a dual-model strategy using IndicTrans2_1B and NLLB_3.3B for low-resource Indic translation. Their system applied cross-iterative back-translation of monolingual data to create high-quality pseudo-parallel corpora and semantic filtering (all-mpnet-base-v2) to enhance domain similarity.

For the Manipuri language, the team BVSLP achieved a higher BLEU score of 4.15 in the primary system and also achieved a higher cosine similarity of 89.60. In contrastive system submission team TranssionMT achieved BLEU score of 4.49 for the en-mni direction. Team BVSLP pipeline integrates a rule-based named entity recognition and translation module prior to NMT training, handling organisation and location names via translation or transliteration from a knowledge base. After preprocessing, byte pair encoding (BPE) was applied to prepare data for Transformer-based NMT training. For the mni-en direction team TranssionMT achieved higher BLEU scores of 13.37,14.86, and cosine similarities of 0.859, 0.860 in both primary and contrastive systems using their dual model strategy.

For the Khasi language, team DoDS-IITPKD achieved a higher BLEU score of 14.20 in primary system and TranssionMT achieved higher BLEU score of 82.56 in contrastive system for the en-kha direction. This team also achieved higher BLEU score of 4.31 in kha-en direction with their primary system submission, while Trans-

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	**primary**	**23.20**	**0.703**	**0.699**	**67.71**	**55.82**	**0.920**
TranssionMT	contrastive1	22.41	0.699	0.705	67.08	55.37	0.918
DoDS-IITPKD	contrastive	21.75	0.690	0.703	65.77	53.77	0.913
DoDS-IITPKD	primary	21.40	0.695	0.701	66.14	54.90	0.918
TranssionMT	contrastive2	20.09	0.709	0.694	67.42	60.76	0.929
RBG-AI	contrastive	15.27	0.626	0.632	60.36	68.50	0.882
DELAB-IIITM	primary	15.02	0.604	0.605	59.37	75.25	0.869
BVSLP	primary	14.91	0.615	0.613	60.29	71.33	0.893
SRIB-NMT	contrastive	12.68	0.004	0.601	57.69	88.43	0.849
AkibaNLP-TUT	primary	12.28	0.539	0.557	55.61	78.24	0.826
MT@HLT-BLR_Amrtita	primary	10.58	0.609	0.539	58.06	148.91	0.875
Transformers	primary	7.63	0.437	0.472	47.36	102.55	0.743
JU-NLP	primary	0.37	0.013	0.022	14.26	116.71	0.039
A3-108	constraint	0.33	0.023	0.021	17.50	286.28	0.077
A3-108	contrastive1	0.33	0.023	0.021	17.50	286.32	0.077
A3-108	contrastive2	0.33	0.023	0.021	17.50	286.21	0.077
A3-108	primary	0.33	0.023	0.021	17.50	286.21	0.077

Table 5: Evaluation results for Assamese → English (as→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive2	67.50	0.772	0.0174	82.46	28.12
TranssionMT	**primary**	**20.97**	**0.470**	**0.0023**	**61.57**	**62.18**
TranssionMT	contrastive1	19.29	0.451	0.0062	60.39	63.86
DoDS-IITPKD	contrastive	17.64	0.422	0.0074	57.71	74.81
DoDS-IITPKD	primary	17.54	0.422	0.0074	57.75	71.17
JU-NLP	primary	16.72	0.412	0.0039	57.22	70.69
DELAB-IIITM	primary	16.11	0.406	0.0030	55.70	68.32
AkibaNLP-TUT	primary	14.03	0.376	0.0132	53.76	74.08
BilbaoMT	contrastive	10.23	0.284	0.0084	43.99	77.84
RBG-AI	contrastive	9.09	0.281	0.0060	45.48	81.73
Transformers	primary	6.92	0.234	0.0010	41.92	89.99
A3-108	constraint	3.03	0.115	0	31.63	108.91
A3-108	contrastive1	3.03	0.114	0	31.30	107.33
A3-108	primary	2.97	0.113	0	31.46	107.35
A3-108	contrastive2	2.93	0.109	0	30.49	104.26
BVSLP	primary	1.81	0.058	0.0030	27.45	98.66
HopeForBest	contrastive	0.00	0.000	0.0000	1.43	152.04

Table 6: Evaluation results for English → Assamese (en→as) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive1	4.49	0.164	0.0125	44.11	85.96
BVSLP	**primary**	**4.15**	**0.146**	**0.0099**	**41.43**	**89.60**
JU-NLP	primary	4.12	0.155	0.0113	43.87	93.16
RBG-AI	contrastive	4.11	0.144	0.0037	40.64	90.04
TranssionMT	primary	3.66	0.135	0.0100	38.64	92.76
SRIB-NMT	contrastive	3.26	0.000	0.0093	39.52	104.80
DELAB-IIITM	primary	3.15	0.113	0.0087	37.51	132.05
Transformers	primary	2.79	0.099	0.0040	33.41	98.76
BilbaoMT	contrastive	2.75	0.091	0.0096	31.69	90.46

Table 7: Evaluation results for English → Manipuri (en→mni) translation direction

sionMT system achieved higher BLEU score of 24.17 with their contrastive system submission. Team DoDS-IITPKD's primary system fine-tuned NLLB-200-distilled-600M for English–Khasi and English–Mizo, and used IndicTrans2 for Assamese and Bodo. The contrastive system incorporated additional corpora (e.g., PMINDIA, Google SMOL) to expand training data. Both systems employed

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	contrastive1	14.86	0.555	0.576	58.08	76.10	0.8601
TranssionMT	**primary**	**13.37**	**0.554**	**0.565**	**56.71**	**79.31**	**0.8599**
JU-NLP	primary	8.10	0.480	0.495	49.60	100.29	0.7974
RBG-AI	contrastive	7.85	0.431	0.468	48.08	94.56	0.7608
DELAB-IIITM	primary	7.35	0.464	0.479	48.78	103.20	0.7645
AkibaNLP-TUT	primary	5.74	0.328	0.370	41.28	109.95	0.6179
Transformers	primary	4.27	0.291	0.327	36.97	125.99	0.5548
BVSLP	primary	3.06	0.221	0.251	35.61	139.03	0.5671
SRIB-NMT	contrastive	0.34	0.026	0.034	12.64	261.49	0.0685

Table 8: Evaluation results for Manipuri → English (mni→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive2	82.56	0.906	0.915	90.17	11.32
DoDS-IITPKD	contrastive	20.08	0.452	0.534	47.36	59.98
ANVITA	contrastive2	19.43	0.457	0.549	45.93	54.41
ANVITA	contrastive	18.83	0.451	0.543	45.48	55.75
DoDS-IITPKD	**primary**	**14.20**	**0.370**	**0.431**	**39.95**	**87.50**
RBG-AI	contrastive	10.31	0.265	0.344	32.10	77.01
BilbaoMT	contrastive	8.03	0.253	0.352	30.32	73.72
ANVITA	primary	7.34	0.248	0.343	28.34	75.77
A3-108	primary	4.26	0.192	0.255	26.80	96.24
A3-108	contrastive2	4.24	0.188	0.254	26.55	94.71
A3-108	contrastive1	4.23	0.193	0.255	26.76	97.94
SRIB-NMT	contrastive	4.19	0.000	0.227	25.63	113.97
A3-108	constraint	4.10	0.194	0.252	26.90	100.62

Table 9: Evaluation results for English → Khasi(en→kha) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	contrastive2	24.17	0.635	0.686	63.04	52.81	0.879
ANVITA	contrastive	7.44	0.376	0.416	41.85	102.87	0.738
RBG-AI	contrastive	5.64	0.273	0.323	36.36	122.12	0.624
DoDS-IITPKD	contrastive	5.52	0.289	0.349	34.85	113.30	0.644
ANVITA	contrastive2	4.39	0.220	0.284	30.65	123.25	0.551
DoDS-IITPKD	**primary**	**4.31**	**0.239**	**0.293**	**31.33**	**131.86**	**0.579**
ANVITA	primary	1.99	0.106	0.137	20.88	223.26	0.297
A3-108	contrastive2	1.09	0.081	0.114	19.26	171.43	0.243
A3-108	contrastive1	1.06	0.080	0.111	19.46	176.13	0.246
A3-108	primary	1.05	0.079	0.111	19.47	177.43	0.246
A3-108	constraint	1.05	0.081	0.111	19.57	179.16	0.247
SRIB-NMT	contrastive	0.34	0.026	0.034	12.64	261.49	0.069

Table 10: Evaluation results for Khasi → English (kha→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
TranssionMT	contrastive2	37.81	0.660	0.704	69.93	41.80
JU-NLP	**primary**	**15.83**	**0.419**	**0.548**	**52.00**	**69.01**
DoDS-IITPKD	contrastive	14.72	0.407	0.506	48.55	69.49
DoDS-IITPKD	primary	14.26	0.415	0.515	48.51	72.22
SRIB-NMT	contrastive	12.45	0.368	0.509	47.53	78.69
RBG-AI	contrastive	12.44	0.359	0.476	46.83	76.47
BilbaoMT	contrastive	11.06	0.325	0.453	40.83	69.20
DoDS-IITPKD	primary (dup)	10.38	0.537	0.576	55.09	86.84

Table 11: Evaluation results for English → Mizo (en→lus) translation direction

LoRA-based parameter-efficient fine-tuning within the Hugging Face Transformers + PEFT frame-

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
TranssionMT	contrastive2	18.45	0.669	0.684	63.13	61.60	0.915
JU-NLP	**primary**	**12.30**	**0.578**	**0.620**	**58.14**	**78.81**	**0.889**
RBG-AI	contrastive	11.92	0.557	0.588	55.76	79.96	0.871
DoDS-IITPKD	contrastive	11.81	0.544	0.581	55.17	74.39	0.865
DoDS-IITPKD	primary	10.38	0.537	0.576	55.09	86.84	0.874
SRIB-NMT	contrastive	0.007	0.001	0.002	6.12	160.06	0.065

Table 12: Evaluation results for Mizo → English (lus→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
ANVITA	**primary**	**11.59**	**0.414**	**0.511**	**49.85**	74.09	0.786
ANVITA	contrastive2	11.25	0.404	0.513	49.36	**73.79**	0.783
ANVITA	contrastive	11.13	0.416	0.506	48.92	74.17	**0.798**
RBG-AI	contrastive	9.62	0.369	0.387	49.43	93.23	0.624
NLPTng-NITAP	primary	5.42	0.307	0.371	41.37	105.61	0.687
A3-108	primary	1.27	0.086	0.121	23.44	138.23	0.211
A3-108	contrastive_2	1.26	0.083	0.119	23.29	139.45	0.203
A3-108	contrastive_1	1.19	0.081	0.116	22.98	145.27	0.205
A3-108	constraint	1.19	0.081	0.113	23.35	147.92	0.201

Table 13: Evaluation results for Nyishi → English (njz→en) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
ANVITA	**primary**	**6.21**	**0.210**	**0.283**	**34.01**	**81.53**
ANVITA	contrastive	5.92	0.203	0.274	34.08	82.83
BilbaoMT	contrastive	3.92	0.132	0.190	29.38	87.77
NLPTng-NITAP	primary	3.40	0.105	0.180	24.58	92.87
RBG-AI	contrastive	2.45	0.080	0.160	12.57	97.19
A3-108	contrastive_2	1.23	0.049	0.078	20.21	120.46
A3-108	primary	1.19	0.049	0.078	20.37	123.93
A3-108	contrastive_1	1.18	0.050	0.077	20.43	124.40
A3-108	constraint	1.17	0.050	0.077	20.65	127.13

Table 14: Evaluation results for English → Nyishi (en→njz) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
DoDS-IITPKD	contrastive	24.97	0.519	0.169	67.81	51.50
DoDS-IITPKD	**primary**	**24.45**	**0.513**	0.168	**67.71**	**51.84**
JU-NLP	primary	19.71	0.455	**0.169**	62.47	64.97
Transformers	contrastive	19.30	0.452	0.168	67.29	72.92
BilbaoMT	contrastive	10.18	0.283	0.160	46.87	71.09
DPKM	primary	4.38	0.132	0.009	35.50	92.56
BVSLP	primary	1.35	0.040	0.168	17.05	106.11
CITK_MT	primary	0.31	0.019	0.003	7.24	808.91
RBG-AI	contrastive	0.20	0.006	0.027	0.81	131.96

Table 15: Evaluation results for English → Bodo (en→bodo) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
DoDS-IITPKD	contrastive	22.11	0.629	0.688	63.55	52.84	0.897
DoDS-IITPKD	**primary**	**21.68**	**0.627**	**0.679**	**62.95**	**54.29**	**0.888**
Transformers	contrastive	11.83	0.526	0.559	54.38	85.73	0.831
RBG-AI	contrastive	1.40	0.071	0.101	19.45	206.05	0.231

Table 16: Evaluation results for Bodo → English (bodo→en) translation direction

work and applied language-specific tagging during preprocessing, achieving a balance between translation quality and computational efficiency.

For the Mizo language, team JU-NLP achieved

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER
ANVITA	contrastive	6.997	0.300	0.367	38.08	76.26
RBG-AI	contrastive	2.220	0.134	0.204	22.85	103.51
ANVITA	**primary**	**1.756**	**0.107**	**0.168**	**18.58**	**104.04**
BilbaoMT	contrastive	1.417	0.076	0.134	20.08	91.93
ANVITA	contrastive2	0.553	0.041	0.054	13.38	335.55
DPKM	primary	0.179	0.006	0.015	5.60	105.49

Table 17: Evaluation results for English → Kokborok (en→trp) translation direction

Team	Submission	BLEU	METEOR	ROUGE-L	ChrF	TER	Cosine Sim.
ANVITA	contrastive	2.99	0.163	0.218	25.52	117.73	0.487
ANVITA	**primary**	**2.41**	**0.108**	**0.175**	**23.55**	**129.15**	**0.359**
RBG-AI	contrastive	1.59	0.086	0.125	20.00	147.75	0.302
ANVITA	contrastive2	0.79	0.051	0.081	16.46	170.61	0.201

Table 18: Evaluation results for Kokborok → English (trp→en) translation direction

higher BLUE score of 15.83 in primary system submission and TranssionMT achieved BLEU score of 37.81 in contrastive system submissions in en-lus direction. JUNLP also achieved higher BLEU score of 12.30 with cosine similarity of 0.889 in primary mode and TranssionMT achieved higher BLUE score of 18.45 with cosinie similarity of 0.915 contrastive submissions in lus-en direction.The JUNLP team addressed English–Assamese, Mizo, Manipuri, and Bodo translation by fine-tuning NLLB and IndicTrans2 using parameter-efficient methods (LoRA and DORA). Their approach emphasized extensive preprocessing including deduplication, script harmonization, and alignment filtering to enhance data quality.

For the Nyishi language, team ANVITA achieved a higher BLUE scores of 11.59 with cosine similarity of 0.786 in primary submission out of all submissions in njz-en direction and achieved higher BLUE score of 6.21 in en-njz direction. Their models employed transfer learning by fine-tuning public pre-trained architectures such as ByT5-base and NLLB-200-distilled-600M, incorporating selective vocabulary expansion and targeted post-editing. The primary submissions utilized organizer-provided datasets, while the contrastive runs applied data augmentation through back-translation, sentence concatenation, and proprietary crawled resources.

For the Bodo language, team DoDS-IITPKD achieved higher BLEU score of 24.45,24.97 in primary and contrastive modes respectively in en-bodo direction. This team also achieved higher BLEU scores of 21.68, 22.11 in primary and contrastive submissions respectively in bodo-en direction.Team DoDS-IITPKD's primary system fine-tuned NLLB-200-distilled-600M for English–Khasi and English–Mizo, and used IndicTrans2 for Assamese and Bodo. The contrastive system incorporated additional corpora (e.g., PMINDIA, Google SMOL) to expand training data. Both systems employed LoRA-based parameter-efficient fine-tuning techniques.

For the Kokborok language, team ANVITA achieved higher BLEU scores of 2.41 in trp-en and 1.756 in en-trp directions with their primary submissions. This team also presented their higher scores in contrastive submissions also. Their models leveraged transfer learning by fine-tuning public pre-trained architectures such as ByT5-base and NLLB-200-distilled-600M, combined with selective vocabulary expansion and targeted post-editing. Primary submissions used organizer-provided datasets, while contrastive runs employed data augmentation via back-translation, sentence concatenation, and proprietary crawled resources. Additionally, language-specific strategies included leveraging Bodo data for Kokborok and tailoring vocabulary for Khasi.

7 Analysis

The evaluation results across multiple translation directions reveal a competitive landscape with significant variations in performance. A key finding is the direct correlation between the size of the parallel training data and the translation quality, although some notable exceptions exist. The figures (Figure 1, 2, and 3) illustrate these findings, providing a visual context for the observations.

Figure 1 shows the best **primary** BLEU score for

each language direction (the highest BLEU among primary submissions for that direction). Figure 2 visualizes BLEU vs. ChrF for the set of all primary submissions.

Observations

- **Correlation with Data Size:** There is a strong general trend that language pairs with larger parallel datasets, such as Assamese and Mizo, have higher translation scores. Conversely, languages with very limited data, like Kokborok, show the lowest performance. This confirms that data scarcity remains a significant bottleneck for low-resource languages.
- **Outlier Performance in Bodo:** A particularly noteworthy finding is the performance of the Bodo language pair. Despite having a relatively small dataset of only 15,215 sentences, it achieved the highest overall BLEU score of 24.45 for the $en \rightarrow bodo$ translation. This suggests that the quality of the Bodo data, or the highly effective model and training strategies employed by teams like DoDS-IITPKD, compensated for the limited size. This performance highlights that data quality and model optimization can sometimes outweigh the sheer quantity of data.
- **Dominance of Key Teams:** Teams such as TranssionMT and DoDS-IITPKD consistently delivered high-performing models, frequently securing the top spot in the language pairs they participated in.
- **Asymmetry in Translation Direction:** A consistent pattern emerged where the translation quality for one direction of a language pair was notably different from the other. This could be due to differences in data quality for each direction or inherent linguistic challenges in translating into a specific language.
- **Correlation of Metrics:** The Figure 2 scatter plot illustrates a clear positive correlation between BLEU and ChrF scores. This indicates that models that perform well on one metric of translation quality generally also perform well on the other, reinforcing the validity of these metrics as indicators of good performance.

Language-wise Analysis

The key findings for each language pair are given below:

- $Assamese \leftrightarrow English$: With one of the largest datasets (54,000 sentences), this pair yielded strong results. TranssionMT was the top performer in both directions, with BLEU scores of 23.20 for $as \rightarrow en$ and 20.97 for $en \rightarrow as$. The performance here aligns with the substantial training data available.
- $Mizo \leftrightarrow English$: This pair also had a large dataset (50,000 sentences), and the results reflect this. JU-NLP consistently outperformed DoDS-IITPKD, achieving the highest BLEU scores for both $en \rightarrow lus$ (15.83) and $lus \rightarrow en$ (12.30).
- $Khasi \leftrightarrow English$: With 26,000 sentences, the performance was moderate. DoDS-IITPKD excelled in this pair, securing the highest BLEU scores in both $en \rightarrow kha$ (14.20) and $kha \rightarrow en$ (4.31). The significant performance gap between the two directions is a point of interest.
- $Manipuri \leftrightarrow English$: Despite a dataset of 23,687 sentences, the $en \rightarrow mni$ direction proved exceptionally challenging, with the top BLEU score being only 4.15. In the reverse direction ($mni \rightarrow en$), TranssionMT led with a much higher BLEU of 13.37. This disparity is a key finding, suggesting that the complexity of translating English into a tonal, agglutinative language like Manipuri is a significant hurdle.
- $Nyishi \leftrightarrow English$: This language pair had the highest sentence count (60,000), but the lowest token count among the larger datasets, suggesting shorter sentences. ANVITA's performance (11.59 and 6.21) was moderate, indicating that sentence quantity alone is not the sole determinant of success.
- $Bodo \leftrightarrow English$: This language pair is the most striking example of data quality and model effectiveness. Despite a small dataset of only 15,215 sentences, DoDS-IITPKD achieved the highest BLEU score (24.45), demonstrating that high-quality data and strong modeling can overcome the limitations of data size.
- $Kokborok \leftrightarrow English$: With the least amount of data (2,269 sentences), this pair

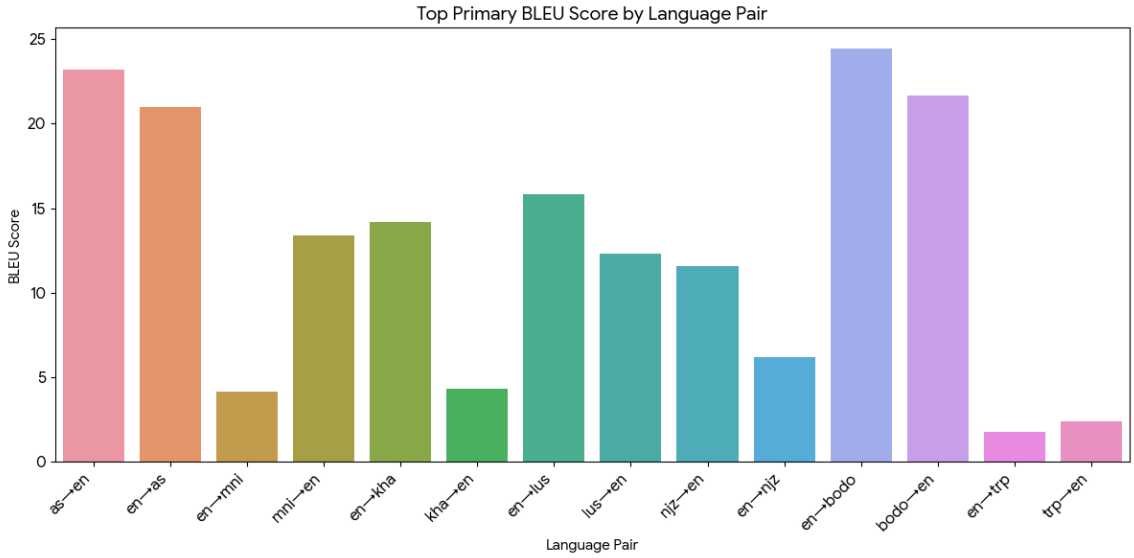


Figure 1: Top BLEU among **primary** submissions per language pair.

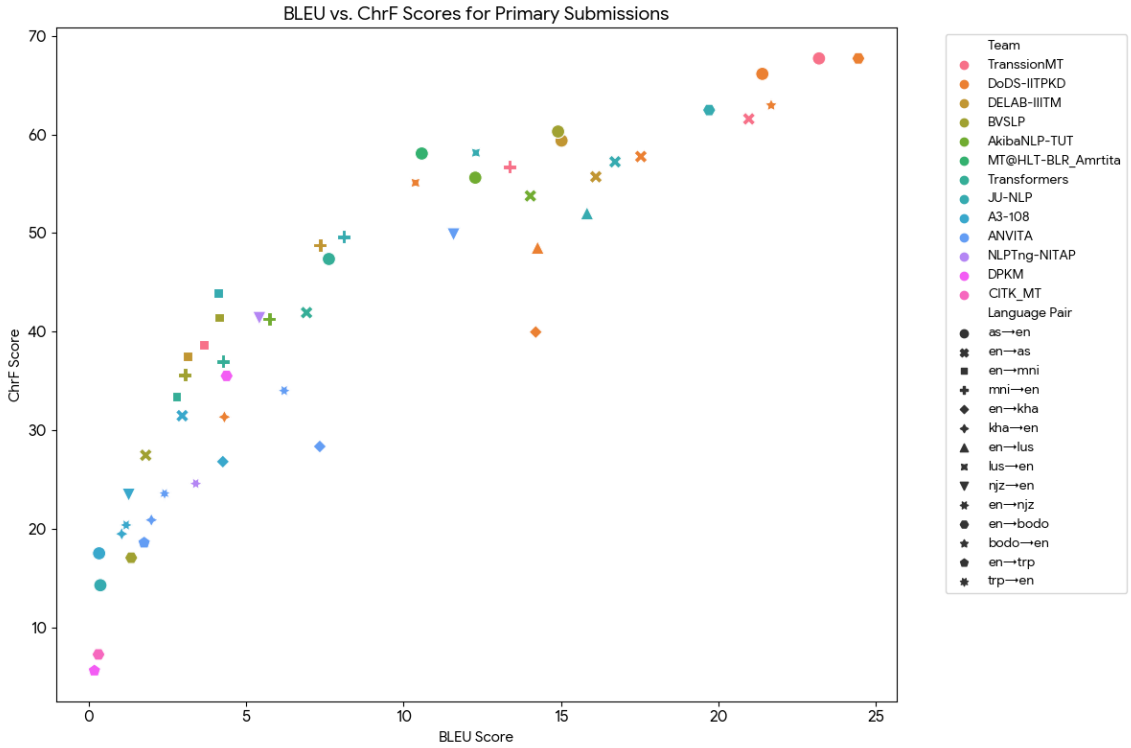


Figure 2: BLEU vs. ChrF for **primary** submissions. Each point is labelled with "team:language".

exhibited the lowest BLEU scores for both directions (1.76 and 2.41), confirming that data scarcity is the primary limiting factor for this language.

Team-wise Analysis

- TranssionMT: This team demonstrated exceptional performance in the Assamese-English and Manipuri-English pairs, consistently ranking at the top. Their models achieved the highest BLEU scores for $as \rightarrow en$ (23.20), $en \rightarrow as$ (20.97), and $mni \rightarrow en$ (13.37), highlighting their strength in these specific languages.

- DoDS-IITPKD: With the highest overall BLEU score of 24.45 for en→bodo, DoDS-IITPKD proved to be a dominant force, espe-

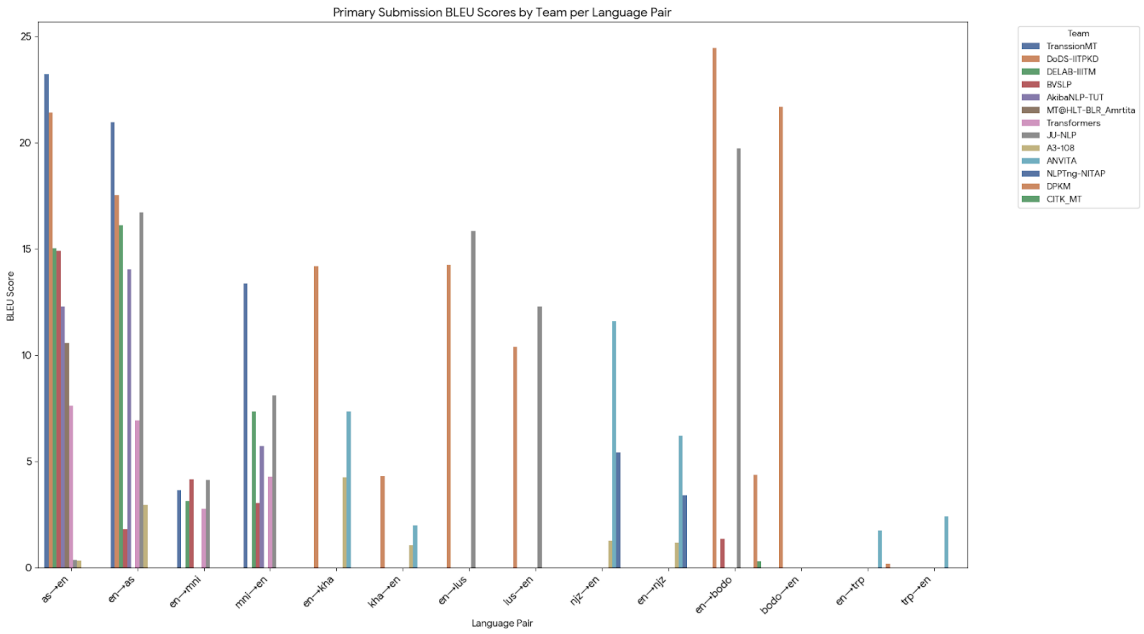


Figure 3: BLEU by team per language pair for **primary** submissions.

cially for the Bodo and Khasi language pairs, where they led in both translation directions.

- JU-NLP: This team's primary submissions were most effective for the Mizo language pairs, where they achieved the highest BLEU scores in both en→lus and lus→en.
- ANVITA: While ANVITA's performance varied, they were the clear leaders in the Nyishi and Kokborok language pairs. Their models, despite the challenges, achieved the best scores for all four directions involving these languages.
- BVSLP: BVSLP's top performance was for the en→mni translation direction, with a BLEU score of 4.15.
- Other Teams: Teams such as DELAB-IIITM, AkibaNLP-TUT, Transformers, and others participated in a number of language pairs, generally achieving lower, though still competitive, scores compared to the top performers.

Categorization of Approaches

Based on the system descriptions, the submitted approaches can be classified into the following methodological categories described in Table19. Note that some teams appear in more than one category when their systems span multiple techniques.

Language-wise Impact and Approach Trends

We further analyze the distribution of techniques across languages for primary submissions in both translation directions (Table 20).

- **>50k pairs** – Teams favored standard Transformer fine-tuning with minor augmentation. Substantial gains were observed from clean fine-tuning alone.
- **20k–30k pairs** – Back-translation and cross-lingual transfer were the most common strategies.
- **<20k pairs** – Parameter-efficient methods and transfer from related languages dominated. Data synthesis played a critical role in achieving competitive performance.

Conclusion

The outcomes of the participating teams in the WMT 2025 translation task have been comprehensively evaluated using a combination of automated metrics and semantic similarity measures. This year's shared task on low-resource Indic language translation utilized the updated **IndicNE-Corp2.0** dataset, which introduced broader domain coverage and incorporated three additional languages Nyishi, Bodo, and Kokborok extending the scope beyond the four language pairs evaluated in 2024. A newly curated test set with higher linguistic and

Category	Teams	Key Characteristics
Phrase-based SMT & Statistical Methods	A3-108	Traditional SMT with BPE, KenLM, deduplication, and focus on vocabulary coverage.
Transformer-based NMT (Standard)	AkibaNLP-TUT, BVSLP, JU-NLP, SRIB-NMT, Transformers	Transformer encoder–decoder architectures (Fairseq, OpenNMT, IndicTrans2, NLLB) with various preprocessing and fine-tuning strategies.
Parameter-efficient Fine-tuning (LoRA, DORA, ReFT)	DoDS-IITPKD, DPKM, JU-NLP, MT@HLT-BLR_Amrita, SRIB-NMT	Efficient adaptation of large multilingual models with minimal computational overhead.
Pretrained Multilingual Models (Zero/Few-shot)	Hope for Best, NLPTng-NITAP, RBG-AI	Direct use of pretrained models (IndicTrans2, mBART50, MADLAD-400) with minimal or targeted adaptation.
Back-translation / Synthetic Data Augmentation	AkibaNLP-TUT, ANVITA, DELAB-IIITM, TranssionMT	Creation of pseudo-parallel data from monolingual corpora to improve low-resource performance.
Transfer Learning & Multilingual Pretraining	ANVITA, BibaoMT, DoDS-IITPKD, Transformers	Leveraging high-resource languages or multilingual corpora to improve target language performance.
Custom Architectures / Specialized Modules	BVSLP (NER module), CITK-MT (GRU + Bahdanau), MT@HLT-BLR_Amrita (ReFT)	Architectures or modules tailored for specific challenges such as named entity handling or fine-grained adaptation.

Table 19: Categorization of submitted systems by methodological approach.

Language (Parallel Data Size)	Common Approaches Seen	Findings
Assamese (54k)	Transformer fine-tuning, direct pretrained model usage, some SMT	Largest resource size in the set; multiple teams reported strong BLEU gains with LoRA fine-tuning.
Mizo (50k)	Transformer fine-tuning + LoRA, back-translation	AkibaNLP-TUT and DoDS-IITPKD achieved consistent gains with monolingual augmentation.
Khasi (26k)	Transfer learning (ByT5, NLLB), BPE-based SMT	SMT still competitive for specific pairs; some systems leveraged Bodo data for transfer.
Manipuri (23.6k)	NLLB fine-tuning, Transformer training, back-translation	Popular among teams due to moderate resource availability.
Nyishi (60k)	SMT + mBART50 fine-tuning	Larger corpus size but fewer participating teams; most relied on transfer learning with prefix tokens.
Bodo (15.2k)	LoRA fine-tuning, ReFT, custom GRU Seq2Seq	Very low-resource; teams adopted parameter-efficient tuning or synthetic data generation.
Kokborok (2.3k)	Transfer learning from related languages, instruction-tuned LLaMA2	Extremely low-resource; innovative data sourcing and vocabulary sharing strategies applied.

Table 20: Language-wise trends in approach adoption for primary submissions in both directions.

structural complexity was also introduced, providing a more rigorous benchmark for system performance. These enhancements are aimed at capturing finer-grained differences in translation quality and reflecting more realistic application scenarios for low-resource Indic languages.

Acknowledgements

We acknowledge the use of linguistic resources and prior descriptive works on Khasi and related languages, which provided valuable background for this study. In particular, we refer to foundational contributions such as (Bareh, 1977), (Grierson, 1928), (Grierson, 1903), (Gurdon, 1904), (Gurdon, 1907), (Nagaraja, 1985), and (Pyrse, 1855).

References

Priyobroto Acharya, Haranath Mondal, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2025. JU-NLP: Improving low-resource indic translation system with efficient lora-based adaptation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

S. K. Acharya. 1971. Languages of khasis. *Mainstream*, May 22:19–26.

James F. Allen. 2003. *Natural language processing*, page 1218–1222. John Wiley and Sons Ltd., GBR.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hamlet Bareh. 1977. *The Language and Literature of Meghalaya*. Indian Institute of Advanced Study, Shimla.

Pramod Chandra Bhattacharya. 1977. *A Descriptive Analysis of the Boro Language*. Department of Publication, Gauhati University, Gauhati, Assam. Originally presented as thesis, Univ. of Gauhati, 1965.

Krishna Boro. 2021. Focus enclitics in bodo. *Linguistics of the Tibeto-Burman Area*, 44(1):75–112.

Census of India. 2011. Language data: Tripura, 2011 census.

Anne Daladier. 2002. Definiteness in amwi: grammaticalization and syntax. *Recherches linguistiques de Vincennes*, 31:61–78. Online edition: mis en ligne le 06 juin 2005; accessed 2025-08-15.

Abhijit Debbarma. 2012. Isolated kokborok vowels recognition. In *Global Trends in Information Systems and Software Applications*, pages 489–493, Berlin, Heidelberg. Springer Berlin Heidelberg.

Khumbar Debbarma, Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. 2012. Morphological analyzer for kokborok. In *Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing*, pages 41–52, Mumbai, India. The COLING 2012 Organizing Committee.

George Abraham Grierson. 1903. *Linguistic Survey of India, Vol II, Part 1*. Banarsidass, Delhi.

George Abraham Grierson. 1928. *Linguistic Survey of India, Vol 11: Mon-Khmer and Siamese-Chinese Family (including Khasi and Tai)*. Motilal Banarsidass, Delhi.

Neha Gupta, Saurabh Salunkhe, Bhagyashree Wagh, and Harish Bapat. 2025. Transformers : Leveraging opennmt and transfer learning for low-resource indian language translation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

P. R. T. Gurdon. 1904. *English Khasi Dictionary*. Mittal Publication, New Delhi.

P. R. T. Gurdon. 1907. *The Khasis*. Macmillan & Co, London.

Loitongbam Gyanendro Singh, Lenin Laitonjam, and Sanasam Ranbir Singh. 2016. Automatic syllabification for Manipuri language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 349–357, Osaka, Japan. The COLING 2016 Organizing Committee.

Barathi Ganesh H and Michal Ptaszynski. 2025. RBG-AI: Benefits of multilingual language models for low-resource languages. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Barry Haddow and Faheem Kirefu. 2020. Pmindia – a collection of parallel corpora of languages of india.

Shoki Hamada, Tomoyoshi Akiba, and Hajime Tsukada. 2025. AkibaNLP-TUT: Injecting language-specific word-level noise for low-resource language translation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, November 15-16,2025 Suzhou, China.

F. Hoque. 2014. Kokborok: A major tribal language of tripura. *IOSR Journal of Humanities and Social Science*.

François Jacquesson. 2003. Kokborok, a short analysis. In *Hukumu, 10th anniversary volume*, pages 109–122. Kokborok Tei Hukumu Mission.

Nisheeth Joshi, Palak Arora, Anju Krishnia, Riya Lonchenpa, and Mahsilenuo Vizo. 2025. BVSLP: Machine translation using linguistic embellishments for indicmt shared task 2025. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, November 15-16,2025 Suzhou, China.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray, and Ajoy Kumar Khan. 2022. Language resource building and English-to-mizo neural machine translation encountering tonal words. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 48–54, Marseille, France. European Language Resources Association.

Ontiwell Khongthaw, G. L. John Salvin, Budde Shrikant Tryambak, Abigairl Nyasha Chigwededza, Dhruvadeep Malkar, and Swapnil Hingmire. 2025. DoDS-IITPKD : Submissions to the wmt25 low-resource indic language translation task. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Deepak Kumar, Laishram Thoibisana Devi, and Asif Ekbal. 2025. DPKM : Tackling low-resource nmt with instruction-tuned llama2: A study on kokborok and bodo. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-English machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022. A domain specific parallel corpus and enhanced english-assamese neural machine translation. *Computación y Sistemas*, 26(4):1669–1687.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shakuntala Mahanta. 2012. Assamese. *Journal of the International Phonetic Association*, 42(2):217–224.

K. S. Nagaraja. 1985. *Khasi: A Descriptive Analysis*. Deccan College, Poona.

K.S. Nagaraja. 2015. Kokborok grammar (an old and rare book). *Exoticindiaart.com*. Published: 2015-05-07.

Dingku Singh Oinam and Navanath Saharia. 2025. DELAB-IIITM : Enhancing low-resource machine translation for manipuri and assamese. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dhrubajyoti Pathak, Sanjib Narzary, Sukumar Nandi, and Bidisha Som. 2025. Part-of-speech tagger for bodo language using deep learning approach. *Natural Language Processing*, 31(2):215–229.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

W. Pyrse. 1855. *Introduction to the Khasi Language*. School Book Society Press, Calcutta.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gargi Roy, Rajesh Kumar, and Kārumūri V. Subbārāo. 2022. Control structures in kokborok: A case of syntactic convergence. *Lingua Posnaniensis*, 63(1):21–52.

Rumphang K. Rynjah and Saralin A. Lyngdoh. 2023. Cross-linguistic comparisons of noun phrase constructions in khasi varieties. *Indian Journal of Language and Linguistics*, 4(2):42–53.

Shaillashree K Sheshadri and Deepa Gupta. 2025. MT@HLT-BLR_Amrita: Bayesian optimization of representation-finetuned adapters for low-resource indic multilingual neural machine translation. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Loitongbam Gyanendro Singh and Sanasam Ranbir Singh. 2017. Word polarity detection using syllable features for manipuri language. In *2017 International Conference on Asian Language Processing (IALP)*, pages 206–209.

J Sivabhavani, Daneshwari Kankawadi, Abhinav Mishra, and Biswajit Paul. 2025. ANVITA : A multi-pronged approach for enhancing machine translation of extremely low-resource indian languages. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, November 15-16,2025 Suzhou, China.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

UNESCO. 2010. Atlas of the world's languages in danger.

Subhash Kumar Wary, Birhang Borgoyary, Akher Uddin Ahmed, Mohanji Prasad Sah, and Apurbalal Senapati. 2025. CITK_MT: An attention-based neural translation system for english to bodo. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

Wikipedia contributors. 2025a. Assamese language. https://en.wikipedia.org/wiki/Assamese_language. Accessed: 2025-08-15.

Wikipedia contributors. 2025b. Boro language (india). https://en.wikipedia.org/wiki/Boro_language_(India). Accessed: 2025-08-15.

Wikipedia contributors. 2025c. Khasi language. https://en.wikipedia.org/wiki/Khasi_language. Accessed: 2025-08-15.

Wikipedia contributors. 2025d. Mizo language. https://en.wikipedia.org/wiki/Mizo_language. Accessed: 2025-08-15.

Wikipedia contributors. 2025e. Nishi language. https://en.wikipedia.org/wiki/Nishi_language. Accessed: 2025-08-15.

Saumitra Yadav and Manish Shrivastava. 2025. A3-108 : A preliminary exploration of phrase-based smt and multi-bpe segmentations for low-resource indian languages. In *Proceedings of the Low-Resource Indic Language Translation Shared Task, 10th Conference on Machine Translation (WMT25), EMNLP*, Suzhou, China.

R. Zothanliana. 2021. *A Study of the Development of Mizo Language in Relation to Word Formation*. Ph.d. thesis, Mizoram University, Aizawl, India.

Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs

Kirill Semenov University of Zurich
Xu Huang Nanjing University
Vilém Zouhar ETH Zurich
Nathaniel Berger Amazon AGI
Dawei Zhu Amazon AGI
Arturo Oncevay Independent
Pinzhen Chen University of Edinburgh & Aveni

Abstract

The WMT25 Terminology Translation Task releases new resources in high-stakes domains and investigates the capabilities of translation systems to accurately and consistently translate specialized terms. This year, we feature new domain and language coverage over previous editions, introducing two distinct tracks: (1) sentence-level translation in the information technology domain for English→German, English→Russian, and English→Spanish, and (2) document-level translation in the finance domain for English↔Traditional Chinese with a document-level one-to-many dictionary. Participants are challenged to translate texts under three modes: no terminology, proper terminology, and random terminology, allowing for a causal analysis of terminology utility. Evaluation combines overall quality, terminology accuracy, and terminology consistency. This shared task attracted broad participation, with 13 teams submitting 20 systems in Track 1 and 4 teams participating in Track 2. The results show that providing proper terminology consistently boosts both overall translation quality and term accuracy, whereas reliance on random terminology yields smaller gains. Despite the near-saturation of sentence-level benchmarks, document-level finance translation still falls short, indicating an urgent need for long-form evaluation and more robust metrics tailored to professional domains.

1 Introduction

Time flies. Since the 2023 edition of the WMT Terminology Translation Task (Semenov et al., 2023), rapid advances in machine translation (MT) and large language models (LLMs) have achieved near-human quality for general-domain translation in several languages (Kocmi et al., 2024, 2025b). Nonetheless, it remains an open question whether these powerful models or techniques can successfully address terminology translation, where the need for accurate and consistent conversion of terminologies poses extra difficulty in addition to general translation quality.

In professional fields like finance, medicine, and law, the correct use of specialized, agreed terms is critical for accuracy and clarity in communications, making terminology translation a research problem of high commercial value (Oncevay et al., 2025b). The field has seen various efforts in modeling (Hasler et al., 2018; Dinu et al., 2019), evaluation (Zouhar et al., 2020; Semenov and Bojar, 2022), and translator tool development (Vargas-Sierra, 2011; Arcan et al., 2017; Lagzdiņš et al., 2022), but recent progress seems to slow down even for high-resourced languages. For example, at the 2024 Conference on Machine Translation, only two papers were dedicated to terminology translation (Kim et al., 2024; Myung et al., 2024).

It is in this context of exciting general domain progress versus modest attention to terminology translation that we organize the WMT25 Terminology Translation Task. The task comes with two primary objectives: (1) to provide an understanding of the current landscape of terminology-aware translation, and (2) to announce and release new, human-annotated datasets to facilitate future research. We organize two tracks covering five translation directions and both sentence- and document-level translation. Particularly, the document track features large document-level one-to-many dictionaries that are more realistic in production. In terms of evaluation, we run multiple metrics that target different facets of terminology conversion, all while taking into account the general quality. Moreover, to estimate the added value (causal effect) of a provided terminology dictionary, systems are evaluated in three translation conditions: without a dictionary, with a proper terminology dictionary, and with a

★ All authors contributed considerably to the design, execution, and presentation of this work. Dawei Zhu's and Nathaniel Berger's work was done outside Amazon. Correspondence to kirill.semenov@uzh.ch. All resources are available at github.com/wmt-conference/wmt25-terminology.

Proceedings of the Tenth Conference on Machine Translation, pages 554–576
November 8-9, 2025

dictionary of random words.

The main findings from the WMT 2025 Terminology Translation Task across submissions are:

★ Most systems involve an LLM in some way. Top systems achieve close to perfect terminology accuracy in the sentence-level track, indicating LLM's suitability for the task, and a saturation in the sentence benchmark. We need to move to document-level benchmarking, which is also more aligned with the practical use.

★ Contrary to the expectations based on previous research, there is little trade-off between general translation ability and terminology accuracy for most participating systems.

★ Incorporating some terminologies always benefits overall translation quality; using proper terms is more useful than random terms for top-performing systems.

2 Task Description

Our task is organized into two tracks, each focused on different domains and input sizes (of both texts and dictionaries), with a unified evaluation protocol. This allows this edition of the terminology shared task to cover wider scope of fields and input types (contrary to (Alam et al., 2021), where only medical domain and sentence-level data were provided), as well as tackling the domains in a more controlled manner (contrary to (Semenov et al., 2023), where every translation direction had its own domain of texts, from NLP abstracts to web novels).

2.1 Tracks and Domains

Track 1: Sentence-level translation

- **Domain:** Information Technology (IT), SAP
- **Translation direction:** English→German, English →Russian, and English→Spanish
- **Setup:** Participants are provided with sentence segments, each with a small terminology dictionary containing only the terms present in the segment, usually 1-2 entries.

Track 2: Document-Level Translation

- **Domain:** Hong Kong finance
- **Translation direction:** English↔Chinese[1]
- **Setup:** Participants are given documents, each accompanied by a large (up to 1K entries) document-level terminology dictionary. English input documents are capped at 2K words; for Chinese→English, the Chinese input is truncated to correspond to the English output of up to 2K words.

  Terminology constraints differ by direction: for English→Chinese, terms are one-to-one mapped, same as in Track 1; for Chinese→ English, terms may be one-to-many mapped, for instance, the target can have both full entity names and acronyms. This track tests terminology accuracy and consistency in long-form translation, reflecting real-world professional needs.

2.2 Terminology Modes

To enable a causal analysis of terminology utility and translation quality, each system is requested to translate our tests under three modes separately:

- **No Terminology (No Term):** The system translates the input text without a terminology dictionary.
- **Proper Terminology (Proper Term):** The system is provided with a dictionary of domain-specific terms; relevant to the input.
- **Random Terminology (Random Term):** The system receives a dictionary of randomly selected source words and their aligned translations from the references (the random pool excludes the proper terms).

The use of Random Term mode allows for measuring if improvements in translation quality stem from incorporating terminologies or from simply seeing part of the correct translation (Zouhar, 2023; Semenov et al., 2023).

3 Data

The data for both tracks[2] is provided in the `jsonl` format, where each instance has the following entries. See Table 1 for an example.

- Text in language 1
- Text in language 2
- Real terminology mapping dictionary (Proper)
- Random word mapping dictionary (Random)
- Dummy empty dictionary (NoTerm)

3.1 Track 1: Sentence-level IT Documentation

Data source. Our sentence-level data was produced by SAP to investigate ambiguous terminology in the IT business domain (Berger et al., 2025).[3] The data originates from their online help

[1] Traditional; henceforth only "Chinese" for brevity.

[2] github.com/wmt-conference/wmt25-terminology/tree/main/ranking/references

[3] github.com/SAP/software-documentation-data-set-for-machine-translation

```
{
    "en": "Open the consumption model containing
         the measures and attributes you want to
         include in your perspective, and click
         the Perspectives tab.",
    "de": "Offnen Sie das Verbrauchsmodell mit
         den Kennzahlen und Attribute, die Sie in
          Ihre Perspektive aufnehmen möchten, un
         wechseln Sie zur Registerkarte
         Perspektiven.",
    "proper": {
        "consumption model": "Verbrauchsmodell"
    },
    "random": {
        "include": "aufnehmen",
        "want": "möchten"
    },
    "noterm": {}
}
```


Table 1: An English→German example from Track 1.

portal. Pages on the help portal are written in English, and translations are produced by post-editing machine translations to ensure proper terminology usage and adherence to corporate style guides.

Terminologies. SAP additionally maintains a one-to-many terminology dictionary across all of its production languages called SAPTerm.[4] Source and target terms were fuzzy-matched in the source and post-edited segments, with additional filtering performed to ensure post-edits made corrections of terminology usage. We make available a terminology dictionary for each segment pair, usually containing one or two entries.

The random terms were retrieved automatically by the following procedure: For every sentence pair (given input and reference translation), the pool of possible random terms is formed by all source sentence words except the terms and the stopwords (based on NLTK2024 stopword lists). Out of this pool, we sample as many words as there are in the corresponding proper dictionary. Then, we prompt ChatGPT to retrieve its translations from the reference sentence. We run an exact match search to ensure that the translation of the word is in the reference. The instruction for ChatGPT is provided in Appendix B.

Test set release. We select the English→ {German, Russian, Spanish} translation directions for the sentence track in our shared task. Per language pair, 1000 instances that contain terminology were sampled, and we split the 1000 segments into development and test sets consisting of 500 instances each. Each instance is accompanied by a terminology dictionary containing entries for that instance only. All test references are only made publicly available after the shared task submission deadline.

3.2 Track 2: Document-Level Finance

Data source. Our document track test data are sourced from the public annual reports on the official website of the Hong Kong Monetary Authority (HKMA), available in English and Traditional Chinese.[5] Each annual report contains multiple chapters, and each chapter is available to download as a standalone PDF file. In this task, we define such a chapter as a *document*. We collect all English and Traditional Chinese annual reports from 2015 to 2024. An annual report yields the same number of chapters (documents) in English and Chinese, and corresponding chapters are parallel to each other.

Document processing. We convert the chapter PDFs into markdown using MinerU (Wang et al., 2024), with table recognition, formula recognition, and optical character recognition disabled. We drop tables and formulae because they consist mostly of numbers without text, and are difficult to translate or evaluate. We then truncate each chapter to 2000 whitespace-delimited English words to keep the documents at a reasonable length for participants. Then, three authors, who are native Chinese speakers fluent in English, manually inspected and processed the markdown files. This includes truncating the Chinese side and re-aligning Chinese and English at the paragraph level in order to fix errors as a result of the automatic processing of the chapter PDFs, which are in a two-column format. As a result, each chapter (document) pair has the same number of lines in both languages.

Terminology extraction and mapping. We extract terms specific to Hong Kong finance from the source and target documents and establish a mapping via two stages. First, we prompt GPT-4.1 with a pair of source and target documents to automatically identify and align terminologies in the two languages, producing a preliminary mapping. Second, two authors independently review the generated mapping and correct it as necessary. Most revisions are removals of relatively generic named entities (e.g., Hong Kong and US dollar) that already have standard translations. The prompt used

[4] sapterm.com

[5] www.hkma.gov.hk/gb_chi/data-publications-and-research/publications/annual-report/

in this first stage is provided in Appendix A for reference. To extract a word mapping for the Random Term mode, we reuse the same technique from the sentence-level Track 1. To better approximate a real-world scenario, we merge the extracted mappings from the chapters (documents) within the same report and generate report-level mappings for both Proper and Random Term modes.

It is worth noting that the HKMA website provides a glossary for Chinese and English separately.[6] However, we did not use it because the Chinese and English lists are independently ordered, are not index-aligned, and contain different numbers of entries. Consequently, it is difficult to construct positional correspondence between entries in the two languages. This can be explored by future work.

Test set release. We release all document-level data we have prepared as the test set for this year's shared task. To avoid temporal bias and to ensure balanced representation of the translation directions, we partitioned the data so that translation direction alternates by year. Reports from odd-numbered years (2015, 2017, 2019, 2021, 2023) are used for English→Chinese tests, whereas those from even-numbered years (2016, 2018, 2020, 2022, 2024) are used for Chinese→English tests. We release each document as a single string, but paragraphs are delimited by `\n\n`, allowing participants to make their own chunking choices. With each test document, a large terminology dictionary is provided, containing mappings for all terminologies in the whole report (i.e., the dictionary is shared between all documents within the report).

4 Metrics and Evaluation

For both tracks, we run reference-based evaluation using gold translations and corresponding terminology dictionaries. We use three types of metrics in our shared task evaluation: overall quality (string match using BLEU and chrF2++; document-level MQM with LLM-as-a-judge), terminology accuracy, and terminology consistency.

This choice of metrics is motivated by two factors: (1) we wish to measure different aspects in translating terminologies, and (2) modern automated metrics can be less robust than a simple string-matching chrF, especially in domains that were not part of the metrics' training data (Lavie et al., 2025; Zouhar et al., 2024a,b).

New this year, we rank submissions based on the Pareto efficiency measured by the quality metric and the terminology. We provide terminology consistency scores as an analysis, and use the document-level AutoMQM scores for a separate ranking in Track 2, as this is not fully empirically validated.

4.1 Overall Translation Quality

BLEU and chrF2++. As an indication of overall quality, we run two reference-based metrics, BLEU (Papineni et al., 2002) and chrF2++ (Popović, 2017), as implemented in sacrebleu (Post, 2018) with default settings.[7,8] In the main paper we report chrF2++ which is tokenization-insensitive, allowing for consistent evaluation of German, Russian, Spanish, English, and Traditional Chinese outputs. Specifically to run the metrics in the document translation track, we treat each entire translation and reference document as a single string (O'Brien et al., 2025).

Doc-level AutoMQM. For document-level translation quality assessment, we use an LLM-as-a-judge: LLMs are prompted to identify translation error spans and assign severity levels, from which the final score is computed. This evaluation method is well interpretable and has been shown to correlate well with human judgment (Kocmi and Federmann, 2023; Freitag et al., 2023, 2024). An extension to the document level is focus-sentence prompting (FSP), which evaluates documents sentence by sentence (Domhan and Zhu, 2025). In FSP, the judge model is provided with the full source and translation as context, along with the specific target sentence to be evaluated. To evaluate document-level translation in Track 2, we use GPT-4o and GPT-5 as judge models, applying the FSP prompt with two modifications: (1) we evaluate three consecutive sentences at a time to improve efficiency; and (2) we provide the judge model with a terminology mapping to better assess translation quality. Details of the judge prompt are provided in Appendix C. Once the model outputs the errors and their severities, we compute the final MQM score for each annual report as a weighted average over the severity levels. We define three categories of severity: minor, major, and critical, with weights

[6] E.g. www.hkma.gov.hk/eng/data-publications-and-research/guide-to-monetary-banking-and-financial-terms/

[7] BLEU|#:1|c:mixed|e:no|tok:{13a,zh}|s:exp|v:2.4.1

[8] chrF2++|#:1|c:mixed|e:yes|nc:6|nw:2|s:no|v:2.4.1

$\text{count}^{\text{src}} \leftarrow 0, \text{count}^{\text{tgt}} \leftarrow 0$
for $\text{src}_i, \text{tgt}_i, d_i \in X$ **do**
  for $\text{term}_j^{\text{src}}, \text{term}_j^{\text{tgt}} \in d_i$ **do**
    if $\text{term}_j^{\text{src}} \in \text{src}_i$ **then**
      $\text{count}^{\text{src}} \leftarrow \text{count}^{\text{src}} + 1$
      if $\text{term}_j^{\text{tgt}} \in \text{tgt}_i$ **then**
        $\text{count}^{\text{tgt}} \leftarrow \text{count}^{\text{tgt}} + 1$
if $\text{count}^{\text{src}} > 0$ **then**
  return $\text{count}^{\text{tgt}}/\text{count}^{\text{src}}$
else
  return 0

Algorithm 1: Terminology Accuracy (Track 1: sentence-level). Input X is a list of source, translation, terminology dictionary triplets $\langle(\text{src}_1, \text{tgt}_1, d_1), \ldots\rangle$.

$A \leftarrow \langle\rangle$ # accuracy for individual terms
for $\text{src}_i, \text{tgt}_i, d_i \in X$ **do**
  for $\text{term}_j^{\text{src}}, \text{Terms}_j^{\text{tgt}} \in d_i$ **do**
    if $\text{term}_j^{\text{src}} \in \text{src}_i$ **then**
      $\text{count}^{\text{src}} \leftarrow \text{src}_i.\textsc{Count}(\text{term}_j^{\text{src}})$
      $\text{count}^{\text{tgt}} \leftarrow 0$
      for $\text{term}_{j,k}^{\text{tgt}} \in \text{Terms}_j^{\text{tgt}}$ **do**
        $\text{count}^{\text{tgt}} \leftarrow \text{count}^{\text{tgt}} + \text{tgt}_i.\textsc{Count}(\text{term}_{j,k}^{\text{tgt}})$
      $A.\textsc{Append}(\textsc{Min}(\frac{\text{count}^{\text{tgt}}}{\text{count}^{\text{src}}}, 1.0))$
if $|A| \neq 0$ **then**
  return $\frac{\sum A}{|A|}$
else
  return 0

Algorithm 2: Terminology Accuracy (Track 2: document-level). Input X is a list of source, translation, terminology dictionary triplets $\langle(\text{src}_1, \text{tgt}_1, d_1), \ldots\rangle$.

of 1, 5, and 10, respectively. For example, an MT system receives an MQM score of 25 if it produces three major errors and one critical error.

4.2 Terminology Accuracy

We also evaluate how accurately translation systems can convert terms based on a given dictionary. We reckon that a source term usually occurs only once in a sentence input, but is more likely to appear multiple times in a document. Thus, we use different implementations for the two tracks as detailed below.

In the sentence track, for each source term appearing in the input text, we check if its corresponding target term appears in the translation, yielding a binary score. The accuracy is then computed as the sum of successful conversions divided by the total number of source words across all input instances. The algorithm is illustrated in Algorithm 1.

At the document level, the accuracy measure is modified to account for: (1) a source word can appear multiple times and thus a target word is expected as many times; (2) potential one-to-many mappings in a dictionary. The metric moves from a binary check to a percentage score. For each source term present in the source document, we calculate a ratio determined by the total number of appearances of all its possible target terms in the translation, divided by the total number of appearances of the source term itself. This ratio for each term is capped at 1 to avoid false positives, and the final document-level terminology accuracy is the average of all individual ratios across all source terms across all documents. The algorithm is shown in Algorithm 2.

The main difficulty of checking terminology accuracy lies in the terminology dictionary usually containing source and target entries in their stem form, but for many languages, we need to capture the inflected forms of the entries too. Hence, when we need to check whether a word is in a segment or count the number of appearances of the word, we always employ a two-pronged matching strategy. First, we run a direct surface-form match between the word and the segment. Second, to account for morphological variations, we check the lowercased lemmatized word against the lowercased lemmatized segment.[9] The final result is the higher value resulting from the two matching strategies—this applies to both binary outcomes or counts.

4.3 Terminology Consistency

We use the framework for the term consistency metric suggested by Semenov and Bojar (2022). The framework allows for automated (and more interpretable than LLM-as-a-judge) evaluation on how consistent the models are when choosing the translation of specialized terms. The modular structure of the framework allows for different levels of strictness in evaluation, so for the current shared task, we focused on two versions of the metric based on term frequency and the dictionary. As illustrated in Algorithm 3, the evaluation consists of the following steps:

- **Preprocessing:** This step requires sentence-level or paragraph-level alignments. Track 1 data already meets this; for Track 2, since the input texts have clear separation between paragraphs (double newline characters), we split the system outputs into segments accordingly.[10]

[9] github.com/stanfordnlp/stanza for lemmatization.

[10] For most systems, this simple preprocessing allowed for consistent alignment. The only exception was that for STITCH outputs, we additionally applied LaBSE embeddings (Feng et al., 2022) to align the split segments.

- **Source term selection:** At this step, we retrieved the subsets of terms present in a given segment. For Track 1, terminology dictionaries already meet the requirement. For Track 2, we filter the document-level dictionary to construct a segment-level dictionary for each segment using a substring match for Chinese and an exact match over lemmatized text for other languages.
- **Term translation alignment:** We then locate the exact part of the output that is a translation of the source term. The most effective way appeared to be few-shot prompting ChatGPT with additional post-processing, with details in Appendix E. We refer to the aligned term translations as "candidates".
- **Pseudo-reference choice:** To estimate the consistency of a system, we need to define "pseudo-references": translations against which we compare candidates. For the main analysis, we choose a frequency criterion: For each source term type, we order the candidate types by their frequency, and define the most frequent one as a pseudo-reference. Notably, this choice is insensitive to the term accuracy: the pseudo-reference may not be the best translation, but it should be used stably over the whole text. In an additional experiment in Appendix F, we also try another pseudo-reference option based on the Proper Term dictionary.
- **Evaluation:** For each term occurrence in each text segment, we check whether the observed translation candidate differs from the pseudo-reference. The final score is formalized as a multi-class accuracy: for each source term type (class), we count the percentage of the candidates matching the pseudo-reference in the submitted texts and run macro-averaging over the class percentages. As a result, we get a score within a range of 0 to 1, which shows the percentage of occurrences of the term translation that are consistent with the chosen pseudo-reference.

for $\text{src}_i, \text{tgt}_i \in X$ **do** # Source term selection & align.
  for $\text{term}_j \in \textsc{SrcTermSelect}(\text{src}_i)$ **do**
    $\text{cand}_j \leftarrow \textsc{Aligner}(\text{src}_i, \text{tgt}_i, \text{term}_j)$
    $\text{CandDict}_{\text{term}_j,\text{cand}_j} \leftarrow \text{CandDict}_{\text{term}_j,\text{cand}_j} + 1$
    $\text{AlgDict}_{i,\text{term}_j} \leftarrow \text{cand}_j$
$\text{PseudRefDict} \leftarrow \{\}$ # Pseudo-reference choice
for $\text{term}_k \in \text{CandDict}$ **do**
  $\text{PseudRefDict}_{\text{term}_k} \leftarrow \textsc{AssignPseudoRef}(\text{term}_k)$
for $\text{src}_i, \text{tgt}_i, \text{AlgDict}_i \in X$ **do** # Scoring
  for $\text{term}_j \in \text{AlgDict}_i$ **do**
    $\text{hit}_{i,j} \leftarrow \mathbb{1}[\text{term}_j = \text{PseudRefDict}_{\text{term}_j}]$
return $\sum_{k \in \text{PseudRefDict}} \frac{\sum \text{hit}_k}{|\text{hit}_k|}$ # Macro-average

Algorithm 3: Terminology Consistency (Track 1: sentence level) with pseudo-reference initialization of the most frequent terms. Input X is a list of source-translation pairs $\langle(\text{src}_1, \text{tgt}_1), \ldots\rangle$.

5 Participants and System Descriptions

This year, apart from our baseline, we see 20 systems in Track 1 and 4 systems in Track 2. Their descriptions are provided below. For an easier navigation over the variety of approaches, we label them with the main features and components of particular submissions, namely:

- models used:
  - NMT NMT model (encoder-decoder)
  - LLM LLM (decoder-only model)
  - multiple models (agents, preprocessing+postprocessing, etc.)
- training data:
  - DatAug data augmentation
  - DatCur data curation (filtering big corpora, enriching training data with annotation, etc.)
- model update techniques:
  - *FT fine-tuning, continuous pre-training, supervised fine-tuning, etc.
  - *PO various types of preference optimization: GRPO, PPO, DPO, etc.
- inference-time strategies:
  - code-switched prompts
  - ICL in-context learning, few-shot prompts, etc.
  - multi-metric decoding (using both general quality and term accuracy for sequence choice)
  - term injection (for NMT models)

o3-term-guide LLM The participant put terminology constraints in the form of explanatory statements and presumably prompted o3 from OpenAI.

DuTerm NMT LLM DatAug *FT This is a two-stage algorithm for terminology translation (Jaswal, 2025). It uses a terminology-aware NMT model fine-tuned from NLLB 3.3B (Costa-Jussà et al., 2022), and prompts GPT-4o for post-editing. To construct the NMT training data, they first extract bilingual terminology dictionaries from WMT25 dev sets, which are then supplemented with terminologies generated by the LLM. Then they use an LLM to synthesize parallel sentences

containing one or more terminologies. Specifically, the terms in both source and target sentences are bounded with special tags for identification. After filtering the training data for quality with COMET-QE and other rules, the NMT undergoes terminology-aware fine-tuning. Given the source, the NMT's translation and term pairs, they prompt GPT-4o to refine the translation for better fluency while keeping the constraints.

Erlendur LLM ICL Ingólfsdóttir et al. (2025) presented an LLM-based translation system using a pipeline approach that combines prompting with modular preprocessing and postprocessing components. In a preparatory stage, the LLM analyzes the source text to extract key terms and idioms, which are then matched with entries from bilingual dictionaries; user-provided glossaries can also be incorporated to enforce consistent terminology. After translation, additional post-processing steps may be applied. For example, a custom seq2seq grammatical error correction model is used to improve Icelandic translations. The system participated in both terminology tracks: for Track 1, it employed its standard pipeline with terminology mapping, while for Track 2, the backbone model was switched from Claude 3.5 to GPT-4.1, as the former refused to translate some test examples.

ISMT-TiU (TiUTerm-V0, TiUTerm-V1) LLM ICL The team submitted two systems, both relying on in-context learning of LLMs, with few-shot examples retrieved with BM25 from the dev set. TiUTerm-V0 is Llama-3-8B (Grattafiori et al., 2024) with 10 in-context examples; TiUTerm-V1 is XGLM-7.5B (Lin et al., 2022) with 12 in-context examples.

Barcelona Supercomputing Center (tower, salamandrata) LLM DatCur *PO Garcia Gilabert et al. (2025) submitted two models: Tower based on Llama2-7B (Touvron et al., 2023) and salamandrata based on Salamandra-7B (Gonzalez-Agirre et al., 2025). They use a novel approach of fine-tuning terminology translation using GRPO (Shao et al., 2024). Specifically, they introduce a terminology adherence reward, which penalizes outputs that do not contain the correct terminology. The training data are based on pseudo-terminology mined heuristically using named entities, noun phrases, and adverbial constructions. The reward during training is joined with a general MT quality reward using a quality estimation model.

IRB-MT (MeGuMa) LLM The submitted system, named MeGuMa, uses LLM agents with terminology-aware translation prompts, in combination with two MT metrics for per-translation-unit solution selection: MetricX (Juraska et al., 2023) and a custom approximation of terminology accuracy which uses an explicit alignment system by Steingrimsson et al. (2023) beyond surface mapping of lemmatized terms. Translation is done in two phases: translation and revision. Models used for translation are taken from three families: Gemma 3 (27B and 12B) by Team et al. (2025), Qwen3 (14B-thinking, 8B-thinking, 14B, 8B) by Yang et al. (2025), and EuroLLM (9B) by (Martins et al., 2025). In the second phase, three models are used to revise all translations: Gemma 3 27B with thinking, Gemma 3 12B with thinking, and Qwen 3 14B with thinking. While Qwen 3 supports thinking natively but not Gemma 3, all of the revision models were prompt-induced to first think and then produce the final solution. The final translation is selected from all of the generated solutions, both initial and revised. The selection criterion is an arithmetic mean of the arithmetic, geometric, and harmonic means of MetricX (Juraska et al., 2024) and terminology accuracy.

CommandA-WMT LLM DatCur *PO Kocmi et al. (2025a) submitted a post-trained version of Command A from Cohere. The data contains a mix of the languages the model was originally trained on, as well as machine translation data in new languages. The data was heavily filtered for quality. This was followed by preference tuning with a bespoke MTExpert dataset for all languages.

BIT LLM *PO Based on Qwen3-8B-Instruct, the participants used the PPO algorithm to perform reinforcement learning according to the terminology accuracy of the model's outputs without using any Dev Data.

Laniqo LLM DatCur *FT ICL Guttmann et al. (2025) use the EuroLLM-9B-Instruct model (Martins et al., 2025) as a foundation. Given an explicit dictionary, the terms in the source sentence in a prompt are substituted with the target language translations of it, creating a code-switched sentence. Additional prompt engineering analysis showed that two-shot prompts were the most efficient. The second modification was fine-tuning on augmented data from OPUS (Tiedemann

and Nygaard, 2004), where the randomly selected source text nouns and verbs were aligned with their translations, and were replaced with them in the same way as the prompts for the LLM. Fine-tuning was conducted with LoRA (Hu et al., 2022). The decoding is done with constraints: inspired by our announced metrics, the Pareto frontier between overall quality and term accuracy, they used epsilon sampling of 100 sentences, followed by multi-dimensional ranking by various QE metrics and term accuracy of a given segment. This approach was named Pareto-Optimal Decoding. The ablation experiments showed that the best scores were achieved by combining a fine-tuned model together with a modified prompt and few-shot examples.

Lingua Custodia (LC-primary, LC-2, LC-3) LLM DatCur *FT *PO Liu et al. (2025) submitted three systems in total: LC-primary, LC-2, and LC-3. They first filtered bilingual data from Common Crawl and WMT25 using LaBSE and applied an unsupervised terminology extraction approach, developed in their 2023 terminology task submission (Liu et al., 2023), to create terminology mappings. They then fine-tuned open-weight and efficient LLMs, Qwen3-4B (Yang et al., 2025) (with thinking mode disabled) and Gemma-3-4b-it (Team et al., 2025). They conducted supervised fine-tuning in the first stage and then applied GRPO in the second stage, using a sentence-level BLEU reward for overall quality and a constraint-following reward for terminology adherence.

CurTermNLLB NMT DatCur *FT Gonzalez-Gomez (2025)'s system is based on LoRA (Hu et al., 2022) fine-tuning of NLLB 200M (Costa-Jussà et al., 2022) on the consumer-grade Apple M3 with an automated pipeline for creating terminology containing data similar to the Track 1 dataset. They select data from OPUS (Tiedemann and Nygaard, 2004), specifically data from the `GNOME`, `KDE4`, and `WikiMatrix` projects. Sentence pairs from this subset of data were embedded with `all-mpnet-base-v2` in Sentence-Transformers (Reimers and Gurevych, 2019), and cosine similarity to Track 1 dev set sentence pairs was computed for filtering together with filtering based on part-of-speech. Source terminology was then aligned to target sentences to create a term dictionary; relevant dictionary entries were provided to the NLLB model as additional input.

UW-BENMT (ContextTerm) NMT DatCur Pong (2025) submitted a system named ContextTerm, a Transformer-based NMT model (roughly Transformer-base size) with terminology-aware data augmentation. The system identifies terminology constraints by selecting source–target alignments whose source words are judged most "important" by the encoder (measured via the norm of their hidden-state vectors) rather than merely low-frequency ones. Training data combined the IT-specific parallel corpora selected with Cross-Entropy Difference filtering and 30k synthetic English sentences generated using Aya-Expanse-8b (Dang et al., 2024), with inline soft constraints applied to 10% of the data.

Multitan (Systran-ft, EuroLLM-ft, MarianMT-ft) LLM NMT DatAug *FT The participants submitted three systems. The general approach was fine-tuning on in-domain data. Specifically, for Systran-ft, the authors used Systran Model Studio Lite to fine-tune Systran's baseline model with augmented in-domain data. For EuroLLM-ft, EuroLLM was updated with in-domain aligned segments and glossary by using LoRA (Hu et al., 2022). For the third system, MarianMT-ft, the team used two fine-tuning strategies: in No Term mode, using the dev set and other in-domain aligned segments; in Proper Term mode, in addition to fine-tuning the model with in-domain segments, they used a glossary for hard-forced training.

TranssionMT *FT This participant used training constraints and post-processing constraints to improve terminology translation accuracy.

STITCH LLM ICL The participants aimed at solving a recently highlighted problem of adding overly large context into prompts. The proposed method is named STITCH, which stands for Structured Terminology Integration for Translation with Context Handling. STITCH makes use of the observation that long-form documents are coarsely aligned on a paragraph-level and injects local terminology context in-flight during generation, while removing already integrated terminology information from the prompt. The approach leads to a task decomposition, allowing the model to perform document-level translation while being guided by local terminology information.

Baseline■ (GPT-4.1-nano) LLM The organizers prepared a baseline approach by querying GPT-4.1-nano (`2025-04-14`) with a long prompt containing

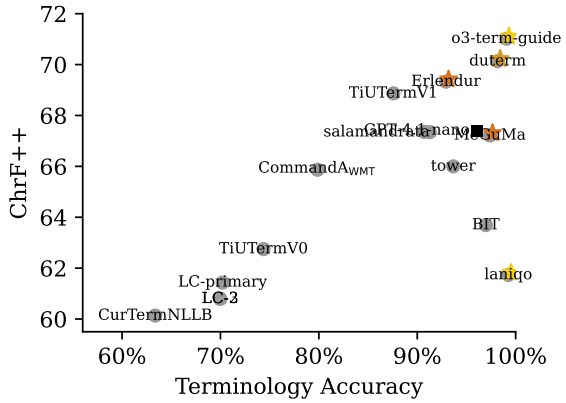


Figure 1: Tradeoff between quality (chrF++) and terminology accuracy in the proper terminology mode, averaged across three directions in Track 1. Top-performing systems are labelled as rank 1★, rank 2★, and rank 3★ according to Pareto optimality.

the input sentence or document and the entire terminology dictionary, when applicable.

6 Shared Task Results

Main results are presented in Table 2 for the sentence-level IT documentation and Table 3 for financial documents. For all three dictionary modes (Proper Term, Random Term, and No Term), the terminology accuracy is *always* measured with respect to the proper terminology dictionary. In addition to chrF++ and terminology accuracy used for system ranking, we also supply terminology consistency, which measures the consistency of the translated terminologies.

6.1 Ranking

We rank all systems in each track by a Pareto efficiency between translation quality (chrF++) and terminology accuracy in the Proper Term mode, where the systems translate real terminologies. For both Track 1 and Track 2, we average chrF++ and term accuracy across all directions for comparison.

Figure 1 visualizes the two-dimensional results for Track 1 systems. In total, there are 5 top-performing systems labelled by ★'s, namely, rank 1 (**o3-term-guide** and **laniqo**), rank 2 (**duterm**), and rank 3 (**Erlendur** and **MeGuMa**), according to Pareto optimality. In Track 2, among the 4 participants, **Erlendur** and **CommandA$_{WMT}$** are at the frontier.

6.2 Translation Quality and Terminology Accuracy

In both tracks, the quality differences between top systems are marginal, as indicated by chrF++ scores, which are usually within 1 chrF++ difference. Terminology handling exhibits sharper contrasts across the two tracks. In the sentence-level Track 1, strong systems achieve very high terminology accuracy of above 97%, implying that state-of-the-art translators can almost perfectly adhere to a few terminology constraints at the sentence level. This also implies that sentence-level terminology translation, as a rather artificial task, lacks difficulty for modern systems. By contrast, the document-level Track 2 exposes the harder challenge, as we observe that accuracy scores drop to the 70–80% range. When many terms need to be translated throughout longer contexts, systems frequently fall short.

Terminology accuracy is not uniform across translation directions. In Track 1, the best systems have similar (and very high) terminology accuracy for the three languages, but a few other systems show some divergence. Our GPT-4.1-nano baseline attains good accuracy when translating into Spanish, but not German or Russian. CommandA$_{WMT}$ and CurTermNLLB show markedly lower accuracy for Russian terminologies compared to Spanish and German. In Track 2, Erlendur delivers better quality and accuracy for English→Chinese, whereas CommandA$_{WMT}$ leads in Chinese→English.

The four systems that entered both tracks (Erlendur, MeGuMa, CommandA$_{WMT}$, and our baseline GPT-4.1-nano) enable us to compare terminology handling under different input lengths and terminology constraint loads. Given longer documents and more terminologies, we expect terminology accuracy to be lower in Track 2. However, CommandA$_{WMT}$ maintains accuracy comparable to, if not better than, its Track 1 results. Moreover, MeGuMa's accuracy is stable across the three languages in Track 1 but shows a dramatic gap of over 30 points between English→Chinese and Chinese→English in Track 2. Nonetheless, we note that the domain and translation direction also changed, which may cause the observed pattern.

Finally, as shown in Figure 1 earlier, there is no clear quality-accuracy tradeoff for many systems; chrF++ and terminology accuracy tend to rise (or drop) together. Outliers such as MeGuMa, tower, BIT, and laniqo lean towards optimizing for termi-

System	Proper, ChrF Avg	Es	De	Ru	Proper, Acc. Avg	Es	De	Ru	Proper, Cons. Avg	Es	De	Ru	Random, ChrF Avg	Es	De	Ru	Random, Acc. Avg	Es	De	Ru	NoTerm, ChrF Avg	Es	De	Ru
o3-term-guide	71.0	75.9	71.6	65.6	99.1	99.1	99.1	99.0	87.7	86.7	86.1	90.4	68.1	72.4	69.4	62.4	49.2	50.7	52.3	44.6	63.6	69.5	64.7	56.6
duterm	70.1	76.1	70.7	63.6	98.2	98.7	98.2	97.6	87.3	86.0	86.3	89.5	66.4	72.1	67.2	59.8	46.6	48.8	48.4	42.4	61.6	67.0	62.6	55.3
Erlendur	69.3	74.8	69.9	63.3	92.9	94.4	93.2	91.2	86.7	83.8	86.3	90.0	66.4	71.6	67.6	59.8	44.4	47.1	47.1	38.9	62.6	68.1	64.0	55.6
TiUTermV1	68.9	77.1	65.7	63.8	87.6	89.4	87.3	86.1	86.7	85.7	85.9	88.5	66.8	74.2	64.4	61.8	54.6	59.2	56.7	47.9	64.4	72.4	61.9	58.9
GPT-4.1-nano■	67.4	72.4	67.4	62.3	90.7	95.2	89.0	88.0	87.5	86.3	86.3	90.0												
salamandrata	67.3	72.0	69.6	60.4	91.3	92.7	91.7	89.4	87.4	87.3	86.4	88.6	64.7	69.3	66.2	58.5	48.2	53.1	48.1	43.4	62.0	67.2	64.0	54.7
MeGuMa	67.2	72.0	67.7	61.9	97.4	97.0	96.3	98.8	88.6	86.9	88.6	90.2	64.5	70.3	64.2	59.0	46.7	53.1	46.4	40.5	58.9	65.2	59.4	52.1
tower	66.0	74.0	65.9	58.1	93.7	95.0	94.8	91.2	88.4	87.6	86.8	90.7	63.8	71.2	63.0	57.1	44.3	48.6	45.7	38.5	60.9	68.6	61.2	53.0
CommandA$_{WMT}$	65.9	70.7	67.6	59.3	79.9	81.9	86.9	70.7	86.6	84.5	87.5	87.8	63.7	68.4	65.0	57.6	45.8	49.3	48.1	40.1	60.7	65.5	62.2	54.4
BIT	63.7	69.8	62.4	58.9	97.0	96.3	98.0	96.7	87.8	86.8	86.9	89.8	65.7	67.2	66.3	63.5	80.5	47.5	97.4	96.5	66.5	69.8	66.3	63.5
TiUTermV0	62.7	69.0	61.0	58.3	74.4	75.2	71.1	76.8	86.4	85.0	85.6	88.6	61.0	68.1	59.1	55.8	49.6	54.2	49.9	44.8	60.2	68.0	57.9	54.6
laniqo	61.7	68.5	59.8	56.9	99.3	98.7	99.4	99.6	87.6	85.6	89.3	87.9	60.2	66.3	59.5	54.8	42.7	46.9	43.5	37.7	55.0	60.3	55.5	49.4
LC-primary	61.4	68.9	61.2	54.2	70.2	74.1	70.7	65.8	85.4	83.6	85.8	87.0	61.0	68.1	59.7	55.2	38.6	43.8	37.4	34.6	57.5	65.0	56.9	50.5
LC-2	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	85.8	85.4	85.7	86.2	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	56.9	64.1	56.8	49.9
LC-3	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	86.0	85.6	85.7	86.7	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	56.9	64.1	56.8	49.9
CurTermNLLB	60.1	69.1	60.3	51.0	63.4	76.5	79.0	34.6	88.0	87.5	87.6	88.8	58.8	67.4	58.0	50.8	36.1	44.1	31.7	32.6	55.6	65.6	52.8	48.4
ContexTerm	48.5	53.7	40.2	51.5	72.0	68.5	79.9	67.6	81.9	75.6	85.8	84.4	48.2	52.0	40.7	51.7	24.6	20.5	18.6	34.8	45.7	50.2	37.4	49.4
Systran-ft		71.1				44.1				88.1				71.1				44.1				71.1		
MarianMT-ft		65.6				17.5				54.1				68.9				48.8				68.9		
EuroLLM-ft		63.5				38.9				82.5				63.5				38.9				63.5		
TranssionMT				47.8				33.2				90.1				47.8				33.2				47.8

Table 2: Main results for Track 1: sentence-level IT documentation terminology-informed translation.

System	Proper, ChrF Avg	EnZh	ZhEn	Proper, Acc. Avg	EnZh	ZhEn	Proper, Cons. Avg	EnZh	ZhEn	Random, ChrF Avg	EnZh	ZhEn	Random, Acc. Avg	EnZh	ZhEn	NoTerm, ChrF Avg	EnZh	ZhEn
Erlendur	60.2	46.1	74.2	78.7	85.4	71.9	92.0	91.6	92.3	57.9	41.8	74.0	64.9	60.1	69.6	57.4	40.8	74.0
CommandA$_{WMT}$	59.6	43.6	75.5	83.6	78.9	88.3	91.5	90.1	93.0	56.7	39.8	73.7	58.8	52.1	65.4	54.9	36.9	72.9
MeGuMa	54.3	39.1	69.4	79.5	96.6	62.4	90.8	93.3	88.3	48.4	31.6	65.2	47.7	43.9	51.5	51.0	33.7	68.3
STITCH	53.4	37.5	69.3	72.8	70.9	74.8	87.4	87.2	87.6	49.9	31.1	68.8	46.9	39.5	54.4	47.5	31.8	63.1
GPT-4.1-nano■	47.9	31.6	64.1	54.7	51.6	57.9	81.9	80.3	83.5	46.5	29.1	63.9	43.8	37.6	50.0	46.1	28.6	63.7

Table 3: Main results for Track 2: document-level finance terminology-informed translation.

nology accuracy, at variable costs in chrF++. A possible explanation for that, at least for MeGuMa and laniqo, can be that the multi-metric optimization used by the authors tends to favor the terminology-specific metrics.

6.3 Terminology Consistency

Tables 2 and 3 show that, contrary to the general MT quality and success rate scores, the spreads of the consistency scores in the Proper Term mode are relatively small, ranging from 0.81 to 0.92. This shows that the models are quite stable in choosing the translations of the specific terms. The performance of the models in Track 2 is overall higher than that of Track 1: the score of 0.87 is among the highest for sentence-level translation and the lowest for document-level translation. A possible reason for that can be the contextual dependency of the generated terms: in a document-level setup, a system attends to previously generated text, and it can be more prone to copying already generated sequences, while each occurrence of a term in a sentence-level setup is translated independently. This is indirectly supported by the observation of another version of the consistency metric: with the "first-seen" pseudo-reference choice. The absolute scores in both versions of the metrics, as well as their rankings, behave in a surprisingly similar manner: the absolute scores of the "first-seen" pseudo-reference initialization are stably lower compared to the "most frequent" initialization by 0.02 on average. This suggests that the first translation of the term would tend to be the most frequent over the document.

Another observation is that, in stark contrast to the terminology accuracy, the system scores are relatively robust to different types (and presence) of explicit terminology. Yet, as was noted for the two main metrics, the difference in consistency between the proper terminology and the two other modes becomes more pronounced in the higher-scoring systems. Such a trend, however, has exceptions: while it is true for English to Russian, German (sentence level), and Chinese (document level) sentence pairs, the English to Spanish and Chinese to English outputs do not show much difference over the whole range of the systems.

Finally, we should note that if the pseudo-

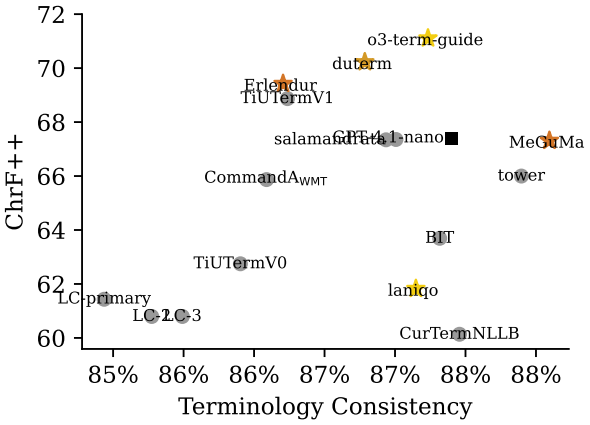


Figure 2: Relationship between quality (chrF++) and terminology consistency in the proper terminology mode, averaged across three directions in Track 1. Top-ranking systems are labelled as in Figure 1.

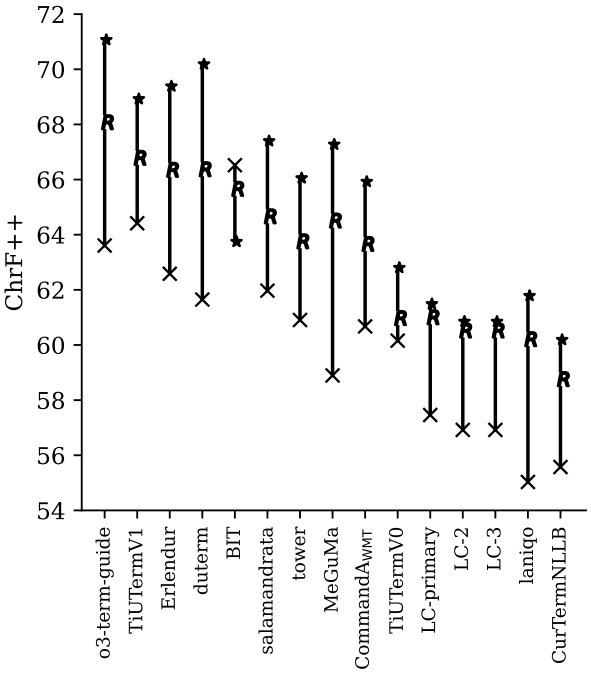


Figure 3: Effect of terminology mode on performance (measured by ChrF++) in Track 1. Legend: $\times$ denotes No Term, R denotes Random Term, and ★ denotes Proper Term.

references are initiated according to the proper terminology dictionaries, the absolute scores drop significantly (resulting in the range of 0.5 to 0.8), and the effect of the terminology mode becomes more pronounced. This is demonstrated in Appendix F, where the scores for each system in Random Term and No Term modes range between 0.2 and 0.4, while the proper terminology lies in a span of 0.5-0.8. Moreover, the difference between modes becomes more pronounced in higher-scoring systems. We conclude that the variant of the metric with dictionary-based pseudo-reference initialization may be more informative for the task of terminology translation, as it correlates with other metrics better and distinguishes between systems more clearly.

7 Analysis

Apart from reporting general translation quality, terminology accuracy, and terminology consistency, we analyze the terminology incorporation and metrics, hoping to provide insights to the community:

- A causal analysis of the impact of incorporating terminology on translation quality.
- A document-level AutoMQM using LLM-as-a-judge, all while considering a large terminology dictionary for Track 2.
- A study of the correlation between different metrics for Track 1.

7.1 Effect of Terminology Incorporation

We investigate the impact of incorporating terminologies on translation quality. Since providing a proper dictionary adds extra target-side information compared to using no dictionary, we also request participants to translate under a Random Term mode to enable a causal analysis. We plot the chrF++ scores under the three terminology modes on a vertical bar in Figure 3 for each participant in the sentence-level Track 1. This helps us easily inspect the quality gap between systems translating under different modes. It is clear that using a dictionary, either random or proper, consistently helps systems' translation quality. For top-performing systems, the benefit of using proper dictionary entries outweighs that of using a random dictionary, but for systems with a lower translation quality, there is no clear difference.

7.2 Doc-Level MQM Results

Table 4 presents the weighted MQM scores for each LLM judge, while Figures 4 and 6 visualize the distribution of error types and severities across the submitted systems in both the No Term and Proper Term settings for GPT-5 and GPT-4o, respectively.

Both LLM judges yield broadly consistent rankings among the top systems, with CommandA$_{WMT}$ and Erlendur generally outperforming MeGuMa and STITCH in terms of overall MQM scores.

	Weighted MQM score					
	GPT-4o			GPT-5		
System	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**
Proper						
Erlendur	75.7	75.4	76.0	29.2	37.5	19.5
CommandA$_{WMT}$	81.2	77.6	84.8	37.7	41.0	34.5
STITCH	174.4	185.6	163.4	57.2	65.8	48.7
MeGuMa	166.8	151.2	182.2	85.1	69.5	100.5
Random						
Erlendur	76.0	81.2	71.0	35.0	49.4	20.8
CommandA$_{WMT}$	85.2	88.9	81.5	56.1	67.7	44.7
STITCH	90.5	89.6	91.3	74.4	83.3	65.8
MeGuMa	217.4	160.2	279.1	186.5	164.4	208.2
NoTerm						
Erlendur	74.0	74.6	73.5	33.1	44.4	22.1
CommandA$_{WMT}$	84.8	86.2	83.5	56.0	62.2	49.9
STITCH	104.9	110.4	99.5	90.4	103.3	77.8
MeGuMa	133.1	125.1	141.0	116.0	117.3	114.7

Table 4: Weighted MQM scores (lower is better, sorted ascending by GPT-5 Proper Avg), averaged over all, EnZh, and ZhEn documents in Track 2. Detailed counts for different error severities are presented in Appendix D.

Lower scores for the leading systems indicate fewer and less severe errors. However, while the overall patterns are similar, the judges diverge in the final ranking of the lower-performing systems in the Proper mode. Notably, GPT-5 tends to be more conservative, flagging fewer errors overall, whereas GPT-4o is stricter in its error identification.

Examining the error type distribution, Figure 4 (GPT-5) shows that most errors are classified as minor or major, with critical errors being relatively rare. The most frequent error types across all severity levels are accuracy, mistranslation, and terminology. The Proper terminology mode consistently reduces the number of terminology-related errors compared to the No Term mode, confirming the utility of providing domain-specific dictionaries.

This trend, however, is not consistently observed with GPT-4o (see Appendix Figure 6). Manual inspection revealed that GPT-4o occasionally produces false positives for terminology mismatch errors, sometimes flagging even exact matches as errors. As a result, we place greater reliance on the GPT-5 results for these outcomes. For future shared tasks, it may be beneficial to combine automated MQM with targeted manual review, or to further refine judge prompts to better accommodate acceptable variation in terminology use.

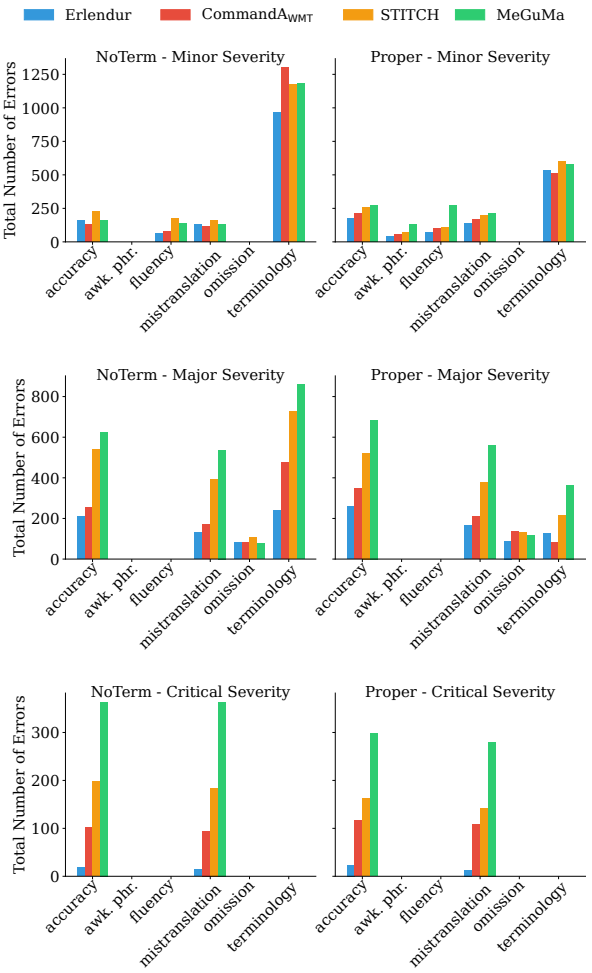


Figure 4: Distribution of error types and severities in No Term and Proper settings using GPT-5 as a judge. Total error counts indicate the number of errors each system made on the complete Task 2 test sets, comprising 10 annual reports across two translation directions. Error types with fewer than 100 occurrences across all severity levels and systems are omitted for clarity. "awk. phr." denotes awkward phrasing.

7.3 Ranking Correlation

One of the reasons for the slow progress in the field of terminology-aware translation is the lack of clarity on the best evaluation protocol(s) due to it being a multifaceted problem. We suggested several types of metrics—general quality, accuracy, and consistency—which differ significantly in the intended focus on the hypothesis and in their implementation. In this section, we analyze how differently or similarly those metrics rank the participating systems alone. This will give insights into the optimal choice of metric(s) in the future.

The comparison of the system ranking by different metrics was conducted using Kendall's τ (Kendall, 1938). We ran the correlation study only

	BLEU	ChrF++	Cons (Freq)	Cons (Dict)
ChrF++	0.81*	–		
Cons (Freq)	0.26	0.39*	–	
Cons (Dict)	0.20	0.36*	0.54*	–
Term Acc	0.22	0.39*	0.49*	0.96*

Figure 5: Kendall's τ correlation between various metrics used in the analysis, Track 1. Comparisons with $p < 0.05$ are marked with *.

for Track 1, since for Track 2, there are only 4 data points, so the τ scores will never surpass the thresholds of statistical significance. We visualize the correlation scores in Figure 5.

We can see two areas of high correlation. The first is well-known, between BLEU and ChrF++. Moreover, terminology accuracy and terminology consistency with dictionary-based pseudo-references display a near-perfect correlation. The reasons for this are clear: both metrics rely heavily on terminology dictionaries; therefore, they are aimed at finding and grouping the proper term occurrences in the system outputs. A surface-level (although lemmatized) match of translated terms is a good approximation of a more tailored system of term alignment between the input and output sequences. Therefore, considering the computational cost of running terminology consistency, future research could rely more on terminology accuracy, as it does not require any external alignment method.

The rest of the metrics show considerably weak positive correlation, in descending order: different types of consistency against term accuracy; and chrF++ against consistency and term accuracy. Although the correlation between the general quality and term-specific metrics is statistically significant, we still conclude that, as shown in Figures 1 and 2, the three types of metrics show different trends. For example, there is a 10 chrF++ gap between the best and second best system in terminology accuracy (Figure 1); top-ranking systems by chrF++ and terminology accuracy do not achieve the best terminology consistency. Therefore, using term-specific metrics is an important aspect of the evaluation of terminology-aware machine translation.

8 Related Work

Past shared tasks. Alam et al. (2021) introduced the first WMT shared task on MT using terminologies, focusing on the medical domain (including COVID-19 terminology) across five language pairs: English to French, Chinese, Russian, and Korean, as well as Czech to German. This pioneering effort established the foundation for systematic evaluation of terminology translation quality and consistency, with terminologies mined semi-automatically from parallel corpora. Building on this, Semenov et al. (2023) organized the second iteration in 2023, which expanded the range of domains (apart from medical texts, it included CL abstracts and web novels), while narrowing down the scope of translation directions: Chinese↔English, English↔Czech, and German↔English. Similar to the previous edition, their terminologies were mined semi-automatically, and they extended this line of work by contrasting random and proper terminologies. Their findings revealed that while incorporating terminology dictionaries led to improvements in translation quality, incorporating equivalent amounts of information from reference translations yielded similar results, challenging the prevailing assumption about terminologies being the crux of meaning in translation. Complementarily, Conia et al. (2025) organized the SemEval-2025 Task 2 on Entity-Aware Machine Translation, which focused on translating text containing complex named entities such as culture-specific titles, location names, and food names across 10 language pairs, introducing the XC-Translate benchmark with over 50K manually-translated sentences with entities that can deviate significantly from word-to-word translations.

Terminology translation test release. To the best of our knowledge, this shared task is among the few that release a high-quality terminology for translation in high-stakes domains such as IT and finance, with the exception of past shared tasks and a contemporary work (Oncevay et al., 2025a).

9 Conclusions

We now conclude the third iteration of the WMT Terminology Translation Task. In comparison to the 2021 and 2023 editions, this time we featured both sentence and document translation tracks with brand new data and domains. The former track ensured continuity, while the latter approximated real-life use cases better. We introduced an LLM-based document-level AutoMQM and used Pareto optimality to rank participants, but we kept the three

inference modes from 2023 for a causal analysis.

We attracted more than 20 submissions, three times more than the previous edition. The overwhelming majority used LLM-based solutions with different types of training techniques. This goes in line with a general trend in the machine translation field towards LLM-based solutions highlighted by Kocmi et al. (2024). Top-scoring systems in the sentence-level track reached good overall translation quality and nearly perfect term accuracy; the document track remains a more challenging task with respect to both metrics. The term consistency, on the contrary, shows a more stable behavior in both tracks, with overall higher scores for document-level MT. In terms of the inference modes, better systems benefit more from proper terminologies, while lower-scoring systems are less sensitive to dictionaries. Finally, we see high correlations between term-based metrics, but not between them and the overall quality, which highlights the necessity to keep at least one terminology-specific metric for this task.

Outlook. The lessons from the shared task also hint at the possible directions for its future iterations:

- Data: continue with document-level terminology translation evaluation
- Metrics: investigate suitable ranking measures and the trade-off between informativeness and computational costs of term-oriented metrics.
- Human evaluation: run human judgment on terminology translations and analyse its correlation with automatic scores. This, to our knowledge, has not been explored before.
- Language: extend the task to more, especially lower-resourced languages, while preventing contamination.

We are open to collaborations, and we especially welcome resources that can be used towards test sets or human evaluation. Stay tuned!

Limitations

The sentence-level test sets have been used in line with their original translation directions; for the document track, we are unsure of the original translation direction, so one of the two directions has the potential problem of translating translated/post-edited text back to its original language.

In terms of evaluation, while we have used several best metrics we can design, there could be some room for considerations and improvements: 1) document-level AutoMQM, especially with terminologies, has not been validated against human judgment; 2) although our terminology match runs lowercasing and lemmatization before string matching, it may not capture all occurrences of an intended word; and 3) certain correlation exists between metrics, e.g. surface string match and terminology match, so they are not fully orthogonal.

Finally, we used a quality-terminology tradeoff to rank participating systems, but as LLMs are more often deployed in practice, cost-effectiveness has become another important aspect.

Acknowledgments

We thank all participants for their submissions.

The document-level finance data used in this shared task is derived from the publicly available annual reports on the Hong Kong Monetary Authority (HKMA)'s website. We acknowledge HKMA as the source and owner of the reports, and we are grateful for the availability of these materials for research purposes.

Pinzhen Chen is supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546].

Vilém Zouhar gratefully acknowledges the support of the Google PhD Fellowship.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2017. Leveraging bilingual terminology to improve machine translation in a CAT environment. *Natural Language Engineering*, 23(5):763–788.

Nathaniel Berger, Johannes Eschbach-Dymanus, Miriam Exel, Matthias Huck, and Stefan Riezler. 2025. Learning to translate ambiguous terminology by preference optimization on post-edits. *arXiv preprint arXiv:2507.03580*.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. SemEval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tobias Domhan and Dawei Zhu. 2025. Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models. *arXiv preprint arXiv:2505.01761*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, and others. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, and others. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*.

Javier Garcia Gilabert, Carlos Escolano, Xixian Liao, and Maite Melero. 2025. Terminology-Constrained Translation from Monolingual Data using GRPO. In *Proceedings of the Tenth Conference on Machine Translation*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, and others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Mariano Gonzalez-Gomez. 2025. CurTermNLLB: Automatic Data Curation and Terminology-Aware Fine-Tuning of NLLB-600M. In *Proceedings of the Tenth Conference on Machine Translation*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kamil Guttmann, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 Terminology Translation Task: A Multi-Objective Reranking Strategy for Terminology-Aware Translation via Pareto-Optimal Decoding. In *Proceedings of the Tenth Conference on Machine Translation*.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank adaptation of large language models. In *International Conference on Learning Representations*.

Svanhvít Lilja Ingólfsdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Þórðarson, and Þorvaldur Páll Helgason. 2025. Miðeind at WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*.

Akshat Jaswal. 2025. It Takes Two: A Dual Stage Approach for Terminology-Aware Translation. In *Proceedings of the Tenth Conference on Machine Translation*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30:81–93.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, and others. 2025a. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation*.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, and others. 2025b. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*.

Andis Lagzdiņš, Uldis Siliņš, Toms Bergmanis, Mārcis Pinnis, Artūrs Vasiļevskis, and Andrejs Vasiljevs. 2022. Open terminology management and sharing toolkit for federation of terminology databases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Kanojia Diptesh, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, and others. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout, and Raheel Qadar. 2023. Lingua custodia's participation at the WMT 2023 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*.

Jingshu Liu, Mariam Nakhlé, Gaëtan Caillaut, and Raheel Qader. 2025. Lingua Custodia's participation at the WMT 2025 Terminology shared task. In *Proceedings of the Tenth Conference on Machine Translation*.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and others. 2025. EuroLLM-9B: Technical report. *arXiv preprint arXiv:2506.04079*.

Jiyoon Myung, Jihyeon Park, Jungki Son, Kyungro Lee, and Joohyung Han. 2024. Efficient technical term translation: A knowledge distillation approach for parenthetical terminology translation. In *Proceedings of the Ninth Conference on Machine Translation*.

Dayyán O'Brien, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*.

Arturo Oncevay, Elena Kochkina, Keshav Ramani, Toyin Aguda, Simerjot Kaur, and Charese Smiley. 2025a. Translating domain-specific terminology in typologically-diverse languages: A study in tax and financial education. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025b. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the*

Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Benjamin Pong. 2025. UW-BENMT at WMT25 Terminology Translation Task: Contextually Selected Pseudo-Terminology Constraints for Terminology-Aware Neural Machine Translation in the IT Domain. In *Proceedings of the Tenth Conference on Machine Translation*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chelo Vargas-Sierra. 2011. Translation-oriented terminology management and ICTs: Present and future. *Interdisciplinarity and languages: Current Issues in Research, Teaching, Professional Applications and ICT*, pages 45–64.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and others. 2024. MinerU: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024a. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024b. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*.

Vilém Zouhar. 2023. Machine translation that peeks at the reference. Report.

A Automatic Terminology Extraction

We prompt GPT-4.1 (`gpt-4.1-2025-04-14`) to automatically extract terminology from the source text and align it with the terminology used in the reference translation. Our prompt is shown in Table 5. Extraction and mapping are performed at the document level. The mappings of documents from the same annual report are then merged into a single terminology mapping for the entire report. In the data release, each source document is accompanied by this report-level terminology mapping, simulating realistic scenarios in which terminology mappings are predefined for a specific domain or task. In such cases, the predefined terms may or may not appear in the source document requiring translation.

B Automatic Random Terminology Alignment

The random terms were aligned with the help of GPT-4o. First, we randomly sample the words from the sentence that are not in the set of the proper terms, then we prompt GPT-4o with the text demonstrated in Table 7. To avoid hallucinations, we post-check the target sentence on whether it contains a highlighted word.

C Focus-Segments Prompting (FSP)

Focus-Sentence Prompting (FSP) was originally proposed by Domhan and Zhu (2025) as a method for using LLMs as judges in long-form translation evaluation while mitigating length bias, which is the tendency of LLMs to underreport errors when evaluating an entire long translation in a single pass. Their approach evaluates one sentence at a time while still providing the entire source and target documents as context. Although effective, the original FSP is costly because it requires many inference calls. To reduce this cost, we introduce Focus-Segments Prompting, in which a segment of three sentences is evaluated at once. This modification reduces the computational cost of FSP by approximately a factor of three. In our meta-evaluation on the WMT'24 Metrics Shared Task data, Focus-Segments Prompting performed comparably to the original FSP.

Another modification we introduced is adapting FSP to better suit our terminology-focused task. We consider accurate terminology translation a key quality dimension that the LLM judge should evaluate. However, even an LLM judge may not always know the correct translations of certain terms. To address this, we provide the judge with a ground-truth terminology mapping for reference. Recall that our original mapping was report-level and included many terms that might not appear in the segment under evaluation. To avoid unnecessary distraction for the judge, we tailor the mapping so that it only retains terms present in the source segment. Furthermore, the judge is explicitly instructed to evaluate terminology usage. Note that providing the mapping means our metric is not entirely reference-free and may correlate more with other terminology-focused metrics. To study this effect, we also tested a standard FSP prompt without access to the terminology mapping across all submissions and settings. We found that the system rankings remained unchanged with the standard FSP prompt, suggesting that the inclusion of the terminology mapping primarily improves the interpretability and focus of the evaluation without fundamentally altering its outcomes.

Our terminology-aware FSP prompt is presented in Table 6.

D MQM Error Count

The output of the MQM judges using the FSP prompt is a list of errors. Each error is assigned a severity of minor, major, or critical. Table 8 reports the number and severity of errors produced by each submitted system, averaged across all documents in Track 2.

E Automatic Term Alignment in the Output Texts

The initial edition of the consistency metric (Semenov and Bojar, 2022) suggested that for term translation, specialized word alignment methods would be used. However, our preliminary analysis shows that both popular solutions, FastAlign by Dyer et al. (2013) and AwesomeAlign by Dou and Neubig (2021) show a lack of robustness with respect to morphological variation of the words, as well as casing and punctuation. Therefore, we used GPT-4o to retrieve the aligned terms from the system outputs. Our experiments showed that few-shot prompting was helpful for the quality of the term retrieval; therefore, we used 20-shot prompts. An example of the alignment prompt can be found in Table 9. For Track 2, we first split the documents into smaller paragraphs and retrieved the subsets of the terms for each segment, i.e., applied the same

```
TASK:
You are an expert linguist and terminologist.

Your job is to:
1. Analyze the source document and identify all domain-specific terminology and key terms (e.g. technical terms, product names,
named entities, etc.).
2. Find the corresponding translations in the translated document.
3. Output the result as a Python dictionary in the format:
{
    "source_term_1": "translated_term_1",
    "source_term_2": "translated_term_2",
    ...
}

RULES:
- Both source and translated documents are in Markdown format and may include image paths (e.g. ![image](path/to/image.png)) or
links. Ignore such elements.
- Only extract relevant terminology – avoid common words, function words, and markdown/control elements.
- If a translation is ambiguous or missing, set the value to null.
- Follow Python dict syntax strictly.
- Do NOT include explanations or extra text – only output the Python dictionary.

{{'-'*40}}
SOURCE DOCUMENT:
{{ source_document }}
{{'-'*40}}
TRANSLATED DOCUMENT:
{{ translated_document }}
{{'-'*40}}

OUTPUT:
(Please provide only the Python dictionary below)
```

Table 5: Prompt used for automatic terminology extraction in Track 2.

preprocessing schema as described in Section 4.3. To avoid hallucinations, every output is compared to a system output (by simple substring search).

We noticed that, while being able to correctly identify the part of the sentence containing a term translation, GPT-4o tends to return an overly long string (for example, if the ground truth term correspondence for English-Spanish sentence is "predefined"-"predeterminado", GPT, given a sentence "Es el valor predeterminado." would return the phrase "**valor** predeterminado" (lit. "predefined **value**". To overcome this, we used the following post-processing schema: each GPT output is compared against the reference term translation. If the number of words in the aligned term is more than it is in the reference translation, we run AwesomeAlign (Dou and Neubig, 2021) on the sentence pair and retrieve the word mappings of each word. Then, we check if the words selected by GPT (lemmatized) indeed correspond to the (lemmatized) source sentence tokens. If not, we cut these words out and leave only the part that corresponds to the exact term translation.

F Consistency Scores with Dictionary-Defined Pseudo-References

Figures 7 and 8 show the term consistency scores of the submitted systems with respect to pseudoreference initialization based on terminology dictionaries. We see that, firstly, the difference between the proper terminology mode, on one side, and random terminology and no terminology modes, on the other side, is significantly larger than in case of most frequent pseudoreference initialization. We also observe the increasing deltas between the proper terminology mode and two other modes in the best scoring systems.

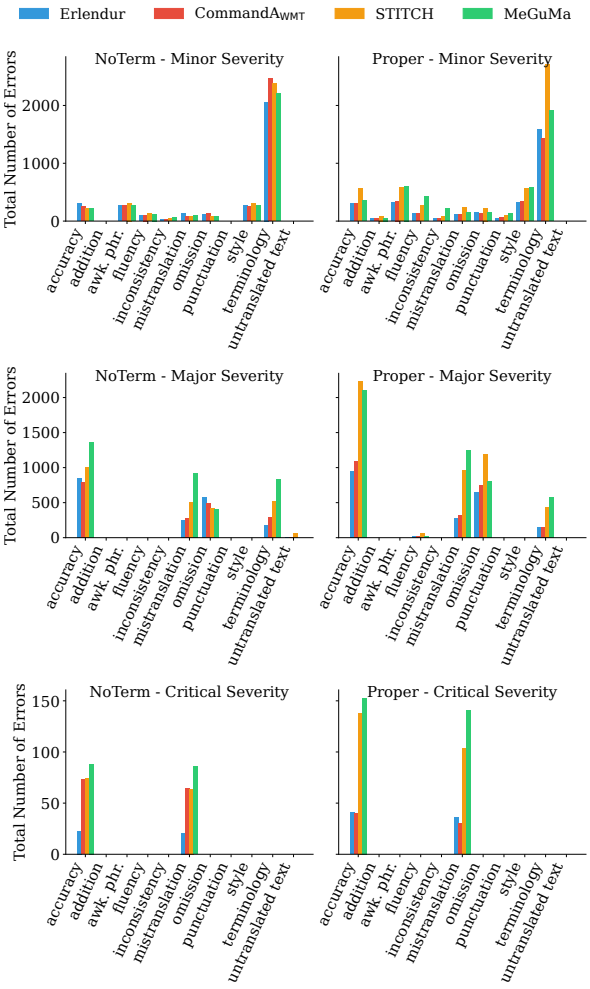


Figure 6: Distribution of error types and severities in No Term and Proper settings using GPT-4o as a judge. Total error counts indicate the number of errors each system made on the complete Task 2 test sets, comprising 10 annual reports across two translation directions. Error types with fewer than 50 occurrences across all severity levels and systems are omitted for clarity. "awk. phr." denotes awkward phrasing.

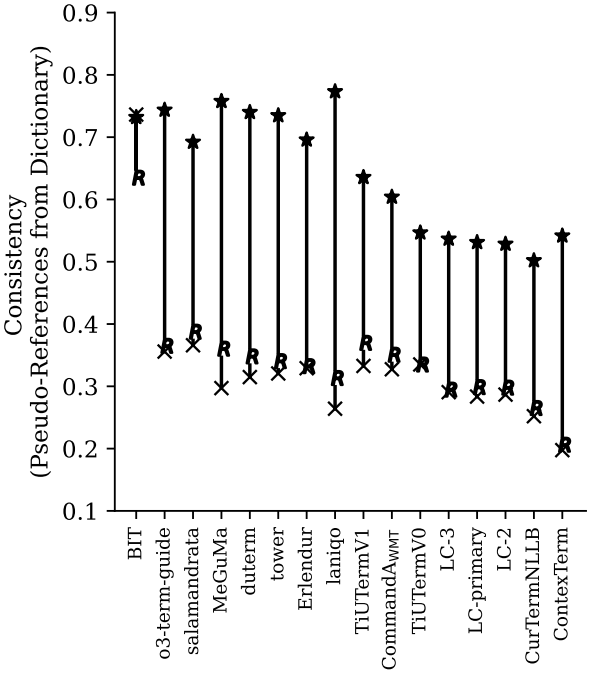


Figure 7: Effect of terminology mode on performance (measured by consistency score with dictionary-defined pseudo-references); Track 1. Legend: × denotes No Term, R denotes Random Term, and ★ denotes Proper Term.

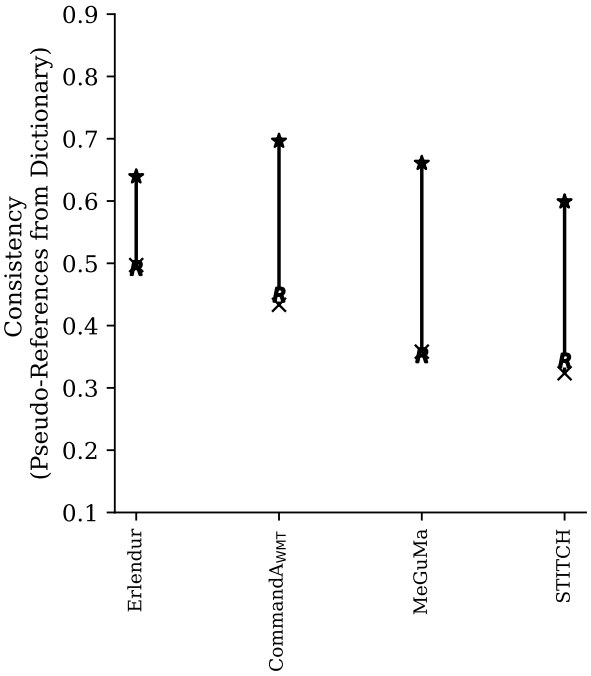


Figure 8: Effect of terminology mode on performance (measured by consistency score with dictionary-defined pseudo-references); Track 2. The legend is identical to that of the Figure 7.

```
You are an annotator for the quality of machine translation. Your task is to identify errors and assess the quality of the
translation using MQM. Based on the source text (in <source></source> tags) and the machine translation (surrounded by
<translation></translation> tags), identify error types in the translation and classify them.

The categories of errors are:
- accuracy (addition, mistranslation, omission, untranslated text, wrong language)
- fluency (character encoding, grammar, inconsistency, punctuation, register, spelling)
- style (awkward phrasing)
- terminology (see subcategories below)
- other

Each error, including omissions or untranslated content, is classified as one of three categories:
- Critical: Errors that make the text incomprehensible or misleading.
- Major: Errors that disrupt flow or distort meaning, but the text is still understandable.
- Minor: Errors that do not affect comprehension but are grammatically, stylistically, or formally incorrect.

The source text must be fully covered, and any omissions should be annotated as errors.
If the error is an omission (missing translation), set "error_span": "" and describe the missing content in "explanation".
Only include spans with errors; exclude correct text.

You will be given a full document and its translations. Only score the sentence in <target_segment></target_segment>, but use the
rest of the document for context. Consistency issues may be flagged per segment, even if similar issues are repeated in other
segments.

Terminology use (if <terminology> is provided; it will be a dictionary and may be empty `{}`):
- Treat <terminology> as a useful reference, not an absolute rule. Prefer its entries when appropriate, but do not penalize
natural, domain-correct alternatives.
- Ignore very minor variations such as capitalization, plural/singular, or the presence/absence of articles (``a,'' ``the'')
unless they clearly change the meaning and/or cause translation errors.
- If an entry in <terminology> seems implausible or clearly incorrect in context, do not enforce it.
- Terminology error subcategories:
  • terminology_mismatch: a correct entry exists in <terminology> for the context, but the translation uses a meaningfully
  different wording that diverges from the provided or established term.
  • terminology_omitted: a source term with a required translation is left untranslated.


Please respond in JSON following this schema:
{
  "type": "object",
  "properties": {
    "errors": {
      "type": "array",
      "items": {
        "type": "object",
        "properties": {
          "error_span": {
            "type": "string",
           "description": "The portion of the translation containing the error. If the error is an omission, use an empty string."
          },
          "explanation": {
            "type": "string",
            "description": "A brief explanation of the error and its impact; reference document context and/or <terminology> only
            if necessary."
          },
          "error_category": {
            "type": "string",
            "enum": ["accuracy", "fluency", "style", "terminology", "other"],
            "description": "The main category of the error."
          },
          "error_type": {
            "type": "string",
            "description": "The specific type of error (e.g., omission, mistranslation, punctuation, terminology_mismatch,
            terminology_omitted)."
          },
          "severity": {
            "type": "string",
            "enum": ["critical", "major", "minor"],
            "description": "Critical: incomprehensible or misleading. Major: distorts meaning but is understandable. Minor: does
            not affect comprehension but is incorrect."
          }
        },
        "required": ["error_span", "explanation", "error_category", "error_type", "severity"]
      }
    },
    "quality_score": {
      "type": "integer",
      "description": "Overall quality score of the translation for the target segment. Use any integer between 0 and 100.
      Guidance: 0 = No meaning preserved, nearly all information is lost. 33 = Some meaning preserved but significant parts are
      missing or garbled; hard to follow. 66 = Most meaning preserved, with only minor grammar/fluency issues; understandable
      overall. 100 = Perfect meaning and grammar, fluent and natural."
    }
  },
  "required": ["errors", "quality_score"]
}

Please score the following input: <input>
<source_language>{{ src_lang }}</source_language>
<source>{{ src }}</source>
<target_language>{{ tgt_lang }}</target_language>
<translation>{{ output_seq }}</translation>
<target_segment>{{ target_segment }}</target_segment>
<terminology>{{ terminology_dict }}</terminology>
</input>

Output requirements:
- Respond with valid JSON only (no text before or after the JSON).
- Produce strings as plain text without Markdown formatting.
```

Table 6: The FSP prompt used to identify MQM errors in the translation.

```
You are a professional translator from {{ source_language }} to {{ target_language }}. You need to help the user with finding the
corresponding words in the {{ target_language }} sentence translated from {{ source_language }}.

Below you are given the {{ source_language }} sentence and its translation, and a list of words to which you need to find the
corresponding words.

You need to return the words and their translations in the form of the Python dictionary, where keys are {{ source_language }}
words and values are their translations, for example, {{ 4-shot example dictionary }}.
DO NOT PRETTIFY THE DICTIONARY, return the raw dictionary in one string.

Source sentence: {{ source sentence }}
Target sentence: {{ target sentence }}
Words that need correspondence: {{ randomly selected words }}
Dictionary of correspondences:
```

Table 7: Prompt for automatically mapping randomly selected words from the source text to the target text.

	MQM Judge: GPT-4o									MQM Judge: GPT-5								
	# Minor			# Major			# Critical			# Minor			# Major			# Critical		
System	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**	**Avg**	**EnZh**	**ZhEn**
Proper																		
Erlendur	21.1	19.0	23.1	10.2	10.9	9.5	0.4	0.2	0.6	7.9	7.9	8.0	3.8	5.2	2.1	0.2	0.3	0.1
CommandA$_{WMT}$	19.8	20.9	18.8	11.6	10.8	12.3	0.4	0.3	0.4	7.8	10.8	5.0	3.9	5.4	2.4	1.1	0.3	1.8
STITCH	37.1	37.4	36.8	25.0	28.1	21.9	1.2	0.8	1.7	9.2	10.3	8.2	6.7	8.6	4.8	1.5	1.3	1.7
MeGuMa	29.6	28.3	30.8	24.7	22.2	27.1	1.4	1.2	1.6	10.8	9.2	12.3	9.5	8.3	10.7	2.7	1.9	3.5
Random																		
Erlendur	24.7	26.0	23.4	10.0	10.8	9.1	0.1	0.1	0.2	11.1	13.9	8.3	4.5	6.7	2.4	0.1	0.2	0.1
CommandA$_{WMT}$	26.4	27.8	25.1	10.6	11.2	10.0	0.6	0.5	0.6	13.0	16.0	10.0	6.8	9.1	4.5	0.9	0.6	1.2
STITCH	25.7	25.2	26.1	11.5	11.8	11.2	0.7	0.5	0.9	14.2	14.6	13.9	9.2	11.2	7.3	1.4	1.3	1.5
MeGuMa	27.4	31.2	23.2	22.3	24.0	20.4	7.9	0.9	15.4	14.9	15.4	14.3	18.9	20.8	17.0	7.7	4.5	10.9
NoTerm																		
Erlendur	24.5	25.8	23.3	9.5	9.5	9.5	0.2	0.1	0.3	11.0	13.3	8.8	4.1	5.7	2.5	0.2	0.2	0.1
CommandA$_{WMT}$	27.9	31.2	24.5	10.1	10.4	9.8	0.7	0.3	1.0	13.8	17.4	10.3	6.6	8.4	4.8	0.9	0.3	1.5
STITCH	27.2	30.4	24.1	14.2	15.3	13.1	0.7	0.4	1.0	14.7	16.4	13.1	11.4	14.1	8.9	1.9	1.6	2.0
MeGuMa	25.4	26.9	24.0	20.0	18.9	21.0	0.8	0.4	1.2	13.7	13.8	13.7	13.4	14.6	12.2	3.5	3.0	4.0

Table 8: Mean number of errors at each severity level (lower is better). Systems are sorted in ascending order, consistent with Table 4. Results are shown overall across 111 documents (Avg) and separately for the EnZh and ZhEn subsets.

```
You are a professional {{ source_language }}-{{ target_language }} translator, teaching the students the course on technical
translation. You are checking a student's translation of a sentence that contains a technical term. You are given an {{
source_language }} term (it can be a word or an expression), a source {{ source_language }} sentence containing this term (it may
be cased differently or contain additional punctuation), and a student's {{ target_language }} translation. You need to find how
the student has translated the term in question in {{ target_language }}, and return only that term.

Important: do not change the translated term anyhow, copy it straight from the sentence! For example, keep the casing and the
grammar form of the translated term as is.

When completing the task, follow the examples below:

        {{ source_language }} sentence: {{ sentence in source language }}
        {{ source_language }} term: {{ source language term }}
        {{ target_language }} translation: {{ reference translation }}
        Translated term: {{ reference term translation }}

                .
                .
                .

        {{ source_language }} sentence: {{ sentence in source language }}
        {{ source_language }} term: {{ source language term }}
        {{ target_language }} translation: {{ reference translation }}
        Translated term:
```

Table 9: Prompt used for automatic terminology alignment. Only one shot of 20 examples was shown explicitly.

System	Proper, ChrF				Proper, Acc.				Proper, Cons.				Random, ChrF				Random, Acc.				Random, Cons.				NoTerm, ChrF				NoTerm, Acc.				NoTerm, Cons.			
	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru	Avg	Es	De	Ru
o3-term-guide	71.0	75.9	71.6	65.6	99.1	99.1	99.1	99.0	87.7	86.7	86.1	90.4	68.1	72.4	69.4	62.4	49.2	50.7	52.3	44.6	88.3	89.1	87.1	88.5	63.6	69.5	64.7	56.6	44.4	46.9	47.5	38.9	89.5	88.8	88.3	91.3
duterm	70.1	76.1	70.7	63.6	98.2	98.7	98.2	97.6	87.3	86.0	86.3	89.5	66.4	72.1	67.2	59.8	46.6	48.8	48.4	42.4	86.6	88.7	84.9	86.3	61.6	67.0	62.6	55.3	42.9	46.9	42.5	39.1	86.9	86.7	86.8	87.0
Erlendur	69.3	74.8	69.9	63.3	92.9	94.4	93.2	91.2	86.7	83.8	86.3	90.0	66.4	71.6	67.6	59.8	44.4	47.1	47.1	38.9	86.2	86.5	84.5	87.5	62.6	68.1	64.0	55.6	42.3	44.9	42.5	39.5	87.1	87.2	86.0	88.0
TiUTermV1	68.9	77.1	65.7	63.8	87.6	89.4	87.3	86.1	86.7	85.7	85.9	88.5	66.8	74.2	64.4	61.8	54.6	59.2	56.7	47.9	85.1	86.4	84.0	84.9	64.4	72.4	61.9	58.9	52.1	54.6	54.1	47.7	85.2	87.9	84.4	83.2
GPT-4.1-nano■	67.4	72.4	67.4	62.3	90.7	95.2	89.0	88.0	87.5	86.3	86.3	90.0																								
salamandrata	67.3	72.0	69.6	60.4	91.3	92.7	91.7	89.4	87.4	87.3	86.4	88.6	64.7	69.3	66.2	58.5	48.2	53.1	48.1	43.4	87.4	87.9	86.3	88.2	62.0	67.2	64.0	54.7	44.4	49.9	42.7	40.7	87.9	88.5	86.9	88.2
MeGuMa	67.2	72.0	67.7	61.9	97.4	97.0	96.3	98.8	88.6	86.9	88.6	90.2	64.5	70.3	64.2	59.0	46.7	53.1	46.4	40.5	87.1	88.4	84.7	88.1	58.9	65.2	59.4	52.1	40.1	46.9	38.5	34.8	86.3	85.6	85.3	87.9
tower	66.0	74.0	65.9	58.1	93.7	95.0	94.8	91.2	88.4	87.6	86.8	90.7	63.8	71.2	63.0	57.1	44.3	48.6	45.7	38.5	87.4	87.7	85.9	88.5	60.9	68.6	61.2	53.0	40.9	46.7	39.6	36.3	88.3	87.9	87.0	90.0
CommandA$_{WMT}$	65.9	70.7	67.6	59.3	79.9	81.9	86.9	70.7	86.6	84.5	87.5	87.8	63.7	68.4	65.0	57.6	45.8	49.3	48.1	40.1	88.3	87.5	86.2	91.3	60.7	65.5	62.2	54.4	43.0	47.3	42.7	38.9	87.7	85.5	86.4	91.3
BIT	63.7	69.8	62.4	58.9	97.0	96.3	98.0	96.7	87.8	86.8	86.9	89.8	65.7	67.2	66.3	63.5	80.5	47.5	97.4	96.5	87.9	87.6	86.8	89.3	66.5	69.8	66.3	63.5	96.7	96.3	97.4	96.5	87.9	87.4	86.9	89.3
TiUTermV0	62.7	69.0	61.0	58.3	74.4	75.2	71.1	76.8	86.4	85.0	85.6	88.6	61.0	68.1	59.1	55.8	49.6	54.2	49.9	44.8	84.9	85.1	84.7	84.9	60.2	68.0	57.9	54.6	49.1	53.6	49.4	44.2	85.2	85.4	84.9	85.4
laniqo	61.7	68.5	59.8	56.9	99.3	98.7	99.4	99.6	87.6	85.6	89.3	87.9	60.2	66.3	59.5	54.8	42.7	46.9	43.5	37.7	82.3	82.9	82.8	81.4	55.0	60.3	55.5	49.4	36.9	41.5	35.2	34.0	82.2	81.0	83.2	82.4
LC-primary	61.4	68.9	61.2	54.2	70.2	74.1	70.7	65.8	85.4	83.6	85.8	87.0	61.0	68.1	59.7	55.2	38.6	43.8	37.4	34.6	85.4	85.8	83.1	87.2	57.5	65.0	56.9	50.5	36.5	41.2	35.5	32.8	84.7	85.3	84.2	84.6
LC-2	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	85.8	85.4	85.7	86.2	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	85.7	86.5	83.7	86.9	56.9	64.1	56.8	49.9	36.3	40.8	35.5	32.6	85.0	85.8	84.3	85.0
LC-3	60.8	67.7	61.0	53.7	70.0	73.6	70.7	65.6	86.0	85.6	85.7	86.7	60.5	67.1	59.5	54.9	38.5	43.4	37.4	34.6	84.9	85.0	83.2	86.5	56.9	64.1	56.8	49.9	36.3	40.8	35.5	32.6	85.3	85.7	84.4	85.7
CurTermNLLB	60.1	69.1	60.3	51.0	63.4	76.5	79.0	34.6	88.0	87.5	87.6	88.8	58.8	67.4	58.0	50.8	36.1	44.1	31.7	32.6	84.1	85.3	82.0	84.9	55.6	65.6	52.8	48.4	34.2	41.7	27.1	33.8	85.7	86.1	84.7	86.2
ContexTerm	48.5	53.7	40.2	51.5	72.0	68.5	79.9	67.6	81.9	75.6	85.8	84.4	48.2	52.0	40.7	51.7	24.6	20.5	18.6	34.8	80.0	75.0	78.3	86.7	45.7	50.2	37.4	49.4	22.4	18.6	13.8	34.8	79.2	72.3	80.8	84.5
Systran-ft		71.1				44.1				88.1				71.1				44.1				88.6				71.1				44.1				88.2		
MarianMT-ft		65.6				17.5				54.1				68.9				48.8				85.1				68.9				48.8				86.4		
EuroLLM-ft		63.5				38.9				82.5				63.5				38.9				83.1				63.5				38.9				82.8		
TranssionMT				47.8				33.2				90.1				47.8				33.2				88.3				47.8				33.2				88.4

Table 10: Extended results for Track 1: sentence-level IT documentation terminology-informed translation. See Table 2 for a subset.

System	Proper, ChrF			Proper, Acc.			Proper, Cons.			Random, ChrF			Random, Acc.			Random, Cons.			NoTerm, ChrF			NoTerm, Acc.			NoTerm, Cons.		
	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn	Avg	EnZh	ZhEn
Erlendur	60.2	46.1	74.2	78.7	85.4	71.9	92.0	91.6	92.3	57.9	41.8	74.0	64.9	60.1	69.6	90.6	89.2	91.9	57.4	40.8	74.0	65.0	60.3	69.6	90.7	89.2	92.1
CommandA$_{WMT}$	59.6	43.6	75.5	83.6	78.9	88.3	91.5	90.1	93.0	56.7	39.8	73.7	58.8	52.1	65.4	90.6	89.1	92.2	54.9	36.9	72.9	56.6	49.1	64.1	90.6	89.0	92.1
MeGuMa	54.3	39.1	69.4	79.5	96.6	62.4	90.8	93.3	88.3	48.4	31.6	65.2	47.7	43.9	51.5	85.8	84.1	87.4	51.0	33.7	68.3	48.3	44.6	51.9	85.1	83.2	87.0
STITCH	53.4	37.5	69.3	72.8	70.9	74.8	87.4	87.2	87.6	49.9	31.1	68.8	46.9	39.5	54.4	84.3	76.9	91.7	47.5	31.8	63.1	44.8	41.2	48.5	84.9	82.7	87.1
GPT-4.1-nano■	47.9	31.6	64.1	54.7	51.6	57.9	81.9	80.3	83.5	46.5	29.1	63.9	43.8	37.6	50.0				46.1	28.6	63.7	43.5	37.2	49.8			

Table 11: Extended results for Track 2: document-level finance terminology-informed translation. See Table 3 for a subset.

Miðeind at WMT25 General Machine Translation Task and Terminology Translation Task

Svanhvít Lilja Ingólfsdóttir, Haukur Páll Jónsson, Kári Steinn Aðalsteinsson, Róbert Fjölnir Birkisson, Sveinbjörn Þórðarson, Þorvaldur Páll Helgason
Miðeind ehf., Reykjavík, Iceland
mideind@mideind.is

Abstract

We present Miðeind's system contribution to two shared tasks at WMT25 – Tenth Conference on Machine Translation: The General Machine Translation Task and the WMT25 Terminology Translation Task. Erlendur is a multilingual LLM-based translation system that employs a multi-stage pipeline approach, with enhancements especially for translations from English to Icelandic. We address translation quality and grammatical accuracy challenges in current LLMs through a hybrid prompt-based approach that can benefit lower-resource language pairs. In a preparatory step, the LLM analyzes the source text and extracts key terms for lookup in an English-Icelandic dictionary. The findings of the analysis and the retrieved dictionary results are then incorporated into the translation prompt. When provided with a custom glossary, the system identifies relevant terms from the glossary and incorporates them into the translation, to ensure consistency in terminology. For longer inputs, the system maintains translation consistency by providing contextual information from preceding text chunks. Lastly, Icelandic target texts are passed through our custom-developed seq2seq language correction model (Ingólfsdóttir et al., 2023), where grammatical errors are corrected. Using this hybrid method, Erlendur delivers high-quality translations, without fine-tuning. Erlendur ranked 3rd-4th overall in the General Machine Translation Task for English-Icelandic translations, achieving the highest rank amongst all systems submitted by WMT25 participants (Kocmi et al., 2025a). Notably, in the WMT25 Terminology Shared Task, Erlendur placed 3rd in Track 1 and took first place in the more demanding Track 2 (Semenov et al., 2025).

1 Introduction

While large language models (LLMs) exhibit strong cross-lingual understanding and can produce high-quality translations from lower-resource languages into major languages like English, gaps in the models' vocabulary and limitations in their grammatical knowledge (Arnett and Bergen, 2025) often become apparent when translating into lower-resource languages (Robinson et al., 2023). Here we describe Erlendur, a multilingual LLM-based translation system designed to address these challenges by enhancing the quality and grammaticality of Icelandic translations. Our main contribution is a hybrid, multi-stage pipeline that combines preparatory text analysis, dictionary lookup, glossary integration, careful prompting, seamless handling of longer texts, and grammatical error correction.

We deployed Erlendur for our submissions to WMT25. In the General Machine Translation Task (unconstrained track)[1] for English-Icelandic translations, Erlendur achieved the highest performance among participating systems for the language pair, ranking 3rd overall behind a human translation (1st) and Gemini 2.5 Pro (2nd). Erlendur marginally outperformed GPT-4.1, though within the margin of statistical significance (Kocmi et al., 2025a). In the WMT25 Terminology Translation Task[2], where participating systems must correctly incorporate glossary terms into their translations, Erlendur placed third in Track 1, which tests injection of glossary terms into short text chunks, and secured first place in Track 2, which tests the scalability of the terminology approach, with much longer glossaries and corpus-level texts (Semenov et al., 2025).

2 System description

Erlendur (see Figure 1) is a translation service provided through Málstaður[3] (Miðeind, 2025), an integrated platform for language technology solutions aimed at Icelandic. Users can access translation

[1] https://www2.statmt.org/wmt25/translation-task.html
[2] https://www2.statmt.org/wmt25/terminology.html
[3] https://malstadur.is

Proceedings of the Tenth Conference on Machine Translation, pages 577–582
November 8-9, 2025

capabilities through Málstaður's editor interface, or through a speech recognition system where they can speak directly and receive translations of the transcribed text. The translation service is also provided commercially through an API.

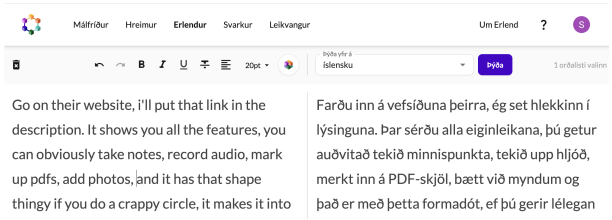


Figure 1: The Erlendur translation interface (in Icelandic) in the Málstaður platform. The user selects the target language; the source language is inferred. "1 orðalisti valinn" indicates that one glossary is selected for use during translation.

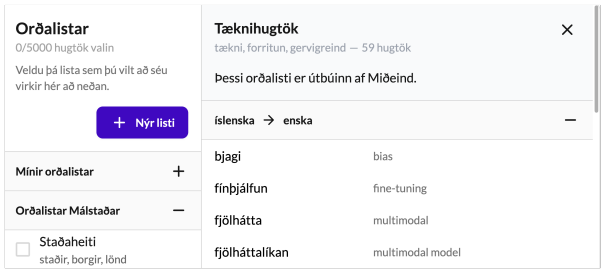


Figure 2: The user interface for the glossary integration in Erlendur (in Icelandic). The user can upload own glossaries or make use of glossaries shared by others in their Málstaður user group.

All the enhancement measures described in this section are intended to improve the translation capabilities beyond a baseline LLM translation. These measures are language-agnostic, unless otherwise stated. The modular nature of the system allows for most of these to be turned on and off. For the WMT25 submissions, all enhancements were used.

2.1 Model selection

While the general language quality of major languages such as English has seen diminishing returns in the latest generation of LLMs, the models still exhibit notable performance gaps when working with many lower-resource languages, and subtle or even significant differences can be noticed between different versions of the same model. We find, for example, that Claude 3.5 Sonnet (Anthropic, 2024) still surpasses other Claude model versions in Icelandic capabilities, even the more recent Claude 4 Sonnet/Opus (Anthropic, 2025). This is supported by Miðeind's Icelandic LLM leaderboard[4] (Miðeind, 2025) results, where Claude 3.5 Sonnet ranks highest of all models in the Claude model family in Icelandic language capabilities.

This motivated our selection of Claude 3.5 Sonnet as the underlying model in Erlendur at the time of system development, and this is the model used for all our WMT25 submissions. The system's API-based architecture makes it straightforward to substitute alternative models as Icelandic language capabilities advance.[5]

2.2 Preparatory analysis and lookup

For translation, the system only needs an input text and a target language; the source language is inferred if none is provided (optional parameters may be supplied). Once the input text is received, it is sent for analysis in a separate pass through the LLM. This is to gain a better understanding of the text, its style, subject domain, and general tone of voice. Key terms and named entities are extracted, and for English-to-Icelandic translations, a list of words is compiled for lookup in Ensk.is, an open-source dictionary[6] (Zoëga and Þórðarson, 2025). The LLM also identifies fixed expressions and idioms that may require special consideration in translation. The results are then prepared for incorporation into the main translation prompt along with the dictionary results provided through the dictionary API. The aim of this preparatory step is to provide the model with richer information on the type of text to translate, and to guide its focus on aspects of the text that require careful translation. This two-pass approach decouples analysis from generation, allowing for more explicit and targeted instructions in the final translation prompt. This step can be further enriched with more versatile tool use, such as dictionaries in other language pairs, or by providing explanations for complex concepts or idioms in the text.

2.3 Glossary integration

An important part of translations in a professional environment is consistent use of terminology. Busi-

[4] https://huggingface.co/spaces/mideind/icelandic-llm-leaderboard

[5] This architecture allowed us, after the WMT25 preliminary results were released (Kocmi et al., 2025b), to easily change the translation model to the high-scoring Gemini 2.5 Pro (https://deepmind.google/models/gemini/pro/), and now, at the time of publication, Gemini 2.5 Pro is the underlying model of Erlendur.

[6] https://ensk.is

nesses often compile glossaries of terms to ensure brand consistency, technical accuracy, and adherence to industry-specific language standards across all translated materials. Erlendur accepts custom glossaries (see Figure 2), where each term can be assigned a subject domain ("finance", "pharmaceutical", etc.) and even a special note on the usage of the term if needed. In the Erlendur editor, the glossary can be provided as a TSV file.

For longer glossaries, it is not feasible to inject them as a whole into the translation prompt, so we need to filter out the terms that appear in the input text. We also need to account for multi-word terms, and properly match those in the input text. Highly-inflected languages present unique challenges for term matching due to morphological variations. We developed a hybrid approach combining fuzzy matching and n-gram analysis, informed by the morphological characteristics of such languages.

The system generates n-grams of lengths 1 through `max_ngram_size` (determined by the longest glossary term) from the input text, then compares each n-gram against glossary terms of the same word count using a string-matching library (RapidFuzz[7]). The algorithm prioritizes longer n-grams first when selecting matches to favor multi-word term matches over shorter partial matches, and uses position tracking to avoid overlapping selections. Fuzzy matching is based on Levenshtein distance between text sequences, with a minimum similarity threshold of 0.75, established through iterative refinement during system development. The goal is to achieve high recall without generating an overly long list of candidate terms. We limit the list to a maximum of 50 terms per text chunk to translate, though this is adjustable, based on the chunk size (see Section 2.4). Once a list of term candidates has been compiled, the terms and their translations are incorporated into the translation prompt as an important resource for the LLM to follow when translating. This curated list of terms likely contains some false positives, but those can simply be ignored by the LLM. This methodology of fuzzy string matching of n-grams allows even inflected terms to be correctly matched, such as when the nominative term "sérstök áætlun" ("specific programme") matches the same term when it appears in the genitive case; "sérstakrar áætlunar". This ensures consistent term use throughout, even for longer documents.

[7] https://github.com/rapidfuzz/RapidFuzz

The glossary functionality is language-agnostic, and since the term matching module was developed to accommodate an inflectional language, it is flexible and should benefit many other morphologically rich languages. While in-house experiments indicate this, formal evaluations remain future work. Of note is that Erlendur placed first in the Terminology Translation Task, Track 2 (see Section 3.2), which tests terminology between Traditional Chinese and English, with Chinese being structurally and morphologically very different from Icelandic. In addition to providing terminological consistency, another benefit of injecting custom glossary terms when translating into lower-resource languages is that they can help fill the vocabulary gaps observed in LLMs for these languages.

2.4 Context-handling and translation

The translation prompt is an information-dense text with clear instructions on how to translate, along with the compiled analysis results, optional dictionary results and relevant glossary terms. An additional information string can be added, with special instructions or information about the text. For longer texts, whose translation might surpass the output token limit of the LLM in question, the system splits the text into fragments or chunks, and translates each chunk separately, while providing a snippet of the previous source text chunk as context to ensure text cohesion.

2.5 Post-processing

LLM-generated texts in Icelandic still contain ungrammatical sentences and made-up words. To remedy this, after translation, we run Icelandic target texts through our in-house grammatical error correction tool, Málfríður (Ingólfsdóttir et al., 2023). This helps catch ungrammatical sentences and correct them, mostly incorrect inflections or unconventional preposition use.

3 WMT submissions

We used Erlendur for both the general and the terminology shared tasks, with the same enhancements. The following sections describe the details of each submission.

3.1 General Machine Translation Task

In our submission, our aim was to use the features already present in Erlendur, without special handling for specific texts in the test set, to demonstrate

the robustness of the system and mimic realistic user behavior. The API offers the option of adding special instructions or information for the task at hand, as mentioned in Section 2.4. This option was used to relay some metadata from the test set, namely the domain, the `doc_id`, and a shortened version of the prompt string provided for each of the four focus domains (*news*, *speech*, *social*, *literary*).

One task-specific instruction was added: A considerable part of the test set data is in the first person (the *speech* domain in particular), and the speaker's gender is not always evident from the text. This calls for some decision-making when translating into Icelandic, to ensure gender agreement in the translation. Icelandic has inherent grammatical gender (masculine, feminine, or neuter), and adjectives change according to gender, so "I'm worried" translates into "Ég er áhyggjufull" (feminine) or "Ég er áhyggjufullur" (masculine), depending on the subject's gender. Instead of inferring the gender from the limited context of the source text, for the *speech* domain, we opted to ask the model to output the standard abbreviated gender notation, "Ég er áhyggjufull(ur)."

Another model-specific limitation is that Claude 3.5 Sonnet cannot produce Icelandic closing quotation marks (“); to remedy this we ask the model to instead output French quotation marks (guillemets, « ») and then we convert them to Icelandic ones („“) in post-processing.

For the translation, we used our standard glossary of place names and organizations in English, and their official Icelandic translations, compiled in-house. This glossary is available for use by Erlendur users.

3.2 Terminology Translation Task

The WMT25 Terminology Translation Task tests the inclusion of a given dictionary of terms when translating, to ensure correct and consistent terminology in specialized domains. This task has two tracks: Track 1 involves translating short text chunks from English into Russian/Spanish/German, and correctly incorporating term translations from a short list of terms that appear in the text. Track 2 better mimics real-life conditions, where the texts are corpus-level and the glossaries are considerably larger. The language pairs are English→Traditional Chinese and Traditional Chinese→English. In both tracks, there are three modes: no terminology, random terminology (the glossary terms are words randomly drawn from input texts) and proper terminology (domain-specific terminology).

In this task, we participated in both Track 1 and Track 2, making use of the native glossary functionality of Erlendur described in Section 2.3. The system's existing capabilities, particularly its efficient term matching and enforcement of terminological consistency, were sufficient to meet the task requirements without further modification. We, however, encountered an unexpected challenge in Track 2. The underlying model, Claude 3.5 Sonnet, refused to consistently generate Chinese translations of the test data, returning a "Content blocked" message. No prompt adjustments we tried could properly bypass this content filter, so we utilized our system's modular architecture to replace the underlying model with GPT 4.1 (OpenAI, 2025), which successfully processed the translations and has solid multilingual capabilities. As the source text itself was innocuous, we concluded that the filter was likely triggered by a policy related to the generation of the target language itself.

This is an example of unforeseen issues that may arise when using external, closed-source models over which the user has no control. It also underscores the value of a flexible system design that permits rapid adaptation, such as swapping the core LLM, to ensure robustness against external constraints.

3.3 WMT results

As briefly mentioned in the introduction, Erlendur ranked 3rd in the WMT25 General Machine Translation Task (Kocmi et al., 2025a) in English-Icelandic translations, with a human translator taking top place and Gemini 2.5 Pro taking second. GPT-4.1 was a close 4th, within the margin of statistical significance. Erlendur thus scored the highest out of all WMT25 participants for this language pair, even with a model that is relatively old and close to being deprecated (Claude Sonnet 3.5). Gemini 2.5 Pro, included for comparison, was the high-scoring model across most languages in the general MT task, also for Icelandic; this has been our cue to replace it as the underlying model in the current version of Erlendur. While Claude 3.5 Sonnet without enhancements was not evaluated in the task, Claude 4 ranked 8-10 for the language pair.

In the Terminology task, Erlendur was the only

system scoring in the top 5 in both Track 1 and Track 2, taking third place in Track 1 and first place in Track 2. This is a clear indicator that our terminology approach, developed with a focus on Icelandic and English, has a solid foundation that works for a range of languages, as different from Icelandic as Russian and Chinese. It also demonstrates its robustness with both longer input texts and its practical value in industry scenarios, where extensive glossaries are commonly used.

4 Conclusion

We have presented Erlendur, a multilingual LLM-based translation system that addresses the challenges of translating into lower-resource languages through a hybrid multi-stage approach. By combining preparatory text analysis, dictionary lookup, glossary integration, and grammatical error correction, the system demonstrates how targeted enhancements can significantly improve translation quality for languages like Icelandic without requiring model fine-tuning. The language-agnostic design of most system components makes this approach applicable to other lower-resource language pairs, while the modular architecture allows for flexible adaptation to different translation scenarios and future model improvements.

Limitations

The translation process of Erlendur is a hybrid pipeline that involves two passes through an LLM, some processing of terms, API lookup and a grammatical correction pass through a separate model, which means that the translation time is longer than in smaller models and simpler solutions. For a faster translation turnaround, each of the preprocessing steps can be included or skipped in the API. The bulk of the overall time, however, is spent on text generation by the model when producing the translation, while the preprocessing steps and dictionary lookups add minimal overhead. Concurrent handling makes time measurements of each step challenging, so while this information would be useful, it has not been reported in this work.

While we hypothesize that our approach, carefully developing a system to translate to and from a morphologically rich language, will benefit other languages of varying morphological complexity, this has not been confirmed with formal experiments.

References

Anthropic. 2024. Claude 3.5 Sonnet Model Card Addendum. https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf.

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf.

Catherine Arnett and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6607–6623, Abu Dhabi, UAE. Association for Computational Linguistics.

Svanhvít Lilja Ingólfsdóttir, Pétur Ragnarsson, Haukur Jónsson, Haukur Símonarson, Vilhjálmur Þorsteinsson, and Vésteinn Snæbjarnarson. 2023. Byte-Level Grammatical Error Correction Using Synthetic and Curated Corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7299–7316, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Miðeind. 2025. Icelandic LLM leaderboard. Accessed: 2025-7-1.

Miðeind. 2025. Málstaður.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025-08-11.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Geir T. Zoëga and Sveinbjörn Þórðarson. 2025. Ensk.is.

A Preliminary Study of AI Agent Model in Machine Translation

Ahrii Kim
AI-Bio Convergence Research Institute
South Korea
ahriikim@gmail.com

trotacodigos/MultiAgentMT.git

Abstract

We present IR_Multi-agentMT, our submission to the WMT25 General Shared Task. The system adopts an AI-agent paradigm implemented through a multi-agent workflow, Prompt Chaining, in combination with RUBRIC-MQM, an automatic MQM-based error annotation metric. Our primary configuration follows the Translate–Postedit–Proofread paradigm, where each stage progressively enhances translation quality. We conduct a preliminary study to investigate (i) the impact of initial translation quality and (ii) the effect of enforcing explicit responses from the Postedit Agent. Our findings highlight the importance of both factors in shaping the overall performance of multi-agent translation systems.

1 Introduction

An AI agent is a computational system that operates autonomously, guided by environmental observations, and often equipped with adaptive learning capabilities (Russell and Norvig, 2010). Large Language Models (LLMs), which have become central to the development of AI agents, demonstrate advanced reasoning, contextual understanding, and flexible workflows across a wide range of tasks, including Machine Translation (MT) (Briva-Iglesias, 2025). Recent work by Briva-Iglesias (2025) explored a multi-agent system composed of four agents—a Translator, Fluency Reviewer, Adequacy Reviewer, and Editor—highlighting its potential for driving future advancements. Inspired by this line of research, we participate in this year's MT track with a multi-agent workflow. Our goal is to leverage smaller models to achieve performance competitive with, or even superior to, larger models, while simultaneously reducing computational costs.

2 Participating Task

The shared task aims to evaluate translation performance across a broad range of languages, domains, genres, and modalities. We participate in the multilingual subtask, which covers 30 target languages, with Czech, English, and Japanese serving as the source languages. Our system is categorized as unconstrained. The source data consists of 29,957 segments, corresponding to 102,060 paragraphs. Our approach performs translation inference on a segment-by-segment basis. In cases where paragraph boundaries are ignored in the original segmentation, we re-split the data into paragraphs and translate them independently.

3 IR_Multi-agentMT

3.1 Design

AI agents enable dynamic workflows through configurable architectures. We adopt the concept of Prompt Chaining, in which the output of each step serves as the input to the next, fostering systematic reasoning and iterative refinement (Briva-Iglesias, 2025). While iterative refinement may theoretically improve translation quality, cost constraints motivate us to employ a unidirectional configuration. Accordingly, we experiment with two workflow variants: Translate–Postedit (TP Workflow) and Translate–Postedit–Proofread (TPP Workflow), both of which are submitted to the competition.

3.2 Translate Agent

The Translate Agent generates target text from the given source using the official prompt provided by the organizers. We use the GPT-4o-mini model to obtain the initial translation.

3.3 Postedit Agent

The Post-edit Agent refines the translation with reference to the source text. We build upon RUBRIC-MQM (Kim, 2025), which, like its counterpart

Proceedings of the Tenth Conference on Machine Translation, pages 583–586
November 8-9, 2025

GEMBA-MQM (Kocmi and Federmann, 2023), identifies MQM-style error categories, severities, and spans. Unlike Gemba-MQM, however, Rubric-MQM reduces biases associated with the MAJOR and MISTRANSLATION labels, improves recognition of NO-ERROR cases, and increases precision in error detection.

We introduce three modifications to the original Rubric-MQM:

Transformation into a post-editor The model is instructed to propose corrected translations for each identified error span.

Severity scale adjustment The original 100-point scale is simplified to a 4-level scale, since severity is not our primary focus. While Kim (2025) stressed that the rubric scheme is essential for model effectiveness, we reduce the rubric complexity in the prompt to streamline usage.

Multilingual configuration While preserving the original in-context examples, we replace one instance from English–German with Japanese–Korean to support broader X–Y translation directions. This modification improves detection of NO-ERROR cases and prevents erroneous corrections of already error-free phrases.

Finally, the model's suggested spans are manually integrated into the sentence, and the revised output is considered the final version.

3.4 Proofread Agent

The Proofread Agent further examines and refines the translation using a Chain-of-Thought prompting strategy (Wei et al., 2022). First, the model identifies potential errors; then it is asked to propose five alternative phrasings that prioritize fluency while maintaining alignment with the source. The best alternative is selected as the final translation. This procedure is designed to resolve awkward expressions introduced during earlier manual edits and to further polish the final output.

4 Model Details

We employ prompt engineering for all agents, using `GPT-4o-mini-2024-07-18` as the primary baseline. The model is configured with a temperature of 1 and a maximum token length of 1024. Although this setup is suboptimal in terms of reproducibility, our iterative pilot studies suggest that these parameters allow the model to explore a wider error space and generate more effective modifications, thereby improving overall translation quality. Future work will focus on developing a more stable and reproducible experimental environment.

5 Experiment

We evaluate our multi-agent pipeline through experiments on the WMT24 English–Spanish dataset (Kocmi et al., 2024). We take a subset of 304 unique source segments with balanced domain distribution (literary, news, social media, and speech) and use 23 translations, summing up to 6,992 segments for analysis.

Our experiments are structured in two parts. First, we obtain initial translations from DeepL (DeepL) and MarianMT (Junczys-Dowmunt et al., 2018), cost-efficient models, and GPT-4o-mini, our baseline model, and compare the final translation quality produced by the multi-agent workflows. We use ChrF++ (Popović, 2017) and COMET (Rei et al., 2020). Second, we analyze translations generated with both the original and the refined versions of Rubric-MQM, with particular attention to cases where (i) no-error labels are produced and (ii) source phrases are erroneously marked as errors.

5.1 Analysis: Initial Translation Quality

Table 1 shows that both the initial and final translation quality are highest when using GPT-4o-mini across the two metrics. Regardless of the initial quality, the general trend in our multi-agent workflow is that translation quality decreases after the Postedit stage and increases again after Proofread, with few exceptions. For instance, the surface-level quality measured by ChrF++ drops from 78.55 to 68.18 with GPT-4o-mini, while COMET scores remain stable, indicating that the revisions primarily involve semantic edits within a similar structural framework. These findings suggest that increasing the extent of semantic-level edits can result in higher overall translation quality.

5.2 Analysis: Response of Postedit Agent

We analyze the erroneous responses produced under the original Rubric-MQM setting. As shown in Table 2, 55.38% of the model outputs are empty, indicating that the system deemed the translation perfect. In 2.8% of the cases, spans were incorrectly labeled as "no-error." Furthermore, in 36.7% of the cases, the model identified source errors, which can disrupt the agent workflow. The revised

Language	Metric	Translate	Postedit	Proofread	Final Δ
DeepL	ChrF++	45.99	38.04	40.62 ↑	-5.37
	COMET	65.72	59.96	59.71	-6.01
MarianMT	ChrF++	60.01	55.92	60.53 ↑	+0.52
	COMET	89.66	87.25	88.40 ↑	-1.26
GPT-4o-mini	ChrF++	78.55	78.93 ↑	68.18	-10.37
	COMET	94.38	91.69	96.05 ↑	+1.67

Table 1: Performance scores of the IR_Multi-agentMT system with different baselines for the initial translation (*Translate*). ↑ indicates gains over the previous stage. The gains from *Translate* to *Proofread* are reported in the *Final* column.

	NaN	No-error	Source error
#	3872	200	2567
%	55.38	2.86	36.71

Table 2: Distribution of erroneous response of Postedit Agent. Raw counts (#) and the percentage (%) are given.

version, however, mitigates these issues through improved prompt engineering.

6 Conclusion

Our experiments on the multi-agent pipeline indicate that higher initial translation quality leads to better final outcomes. Furthermore, enforcing the Postedit Agent to identify more errors is essential for ensuring meaningful revisions within the workflow. In this way, IR_Multi-agentMT demonstrates the potential to achieve translation quality comparable to that of larger models, while operating at roughly half the cost. A more detailed discussion of these findings will be provided in the main system paper.

Acknowledgment

This research was supported by G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317)."

References

Vincent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891*.

DeepL. Deepl translator. https://www.deepl.com/translator. Accessed: 2025-07-05.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *ACL 2025 Industry Track*. Associations for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 1–46, Online / Virtual. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Conference on Machine Translation (WMT), Shared Task Papers, Volume 2*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Alon Lavie Farinha, Luisa Coheur, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.

Stuart Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*, 3rd edition. Prentice Hall, Upper Saddle River, NJ.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Marco Large Translation Model at WMT2025: Transforming Translation Capability in LLMs via Quality-Aware Training and Decoding

Hao Wang, Linlong Xu, Heng Liu, Yangyang Liu, Xiaohu Zhao
Bo Zeng, Longyue Wang, Weihua Luo, Kaifu Zhang

Alibaba International Digital Commerce

https://github.com/AIDC-AI/Marco-MT
https://huggingface.co/AIDC-AI/Marco-MT-Algharb

Abstract

This paper presents the **Marco-MT-Algharb** system, our submission to the WMT2025 General Machine Translation Shared Task from **Alibaba International Digital Commerce** (AIDC). Built on a large language model (LLM) foundation, the system's strong performance stems from novel quality-aware training and decoding techniques: (1) a two-step supervised fine-tuning (SFT) process incorporating data distillation, (2) a two-step reinforcement learning (RL) framework for preference alignment, and (3) a hybrid decoding strategy that integrates word alignment with Minimum Bayes Risk (MBR) re-ranking to improve faithfulness. These approaches jointly ensure high accuracy and robustness across diverse languages and domains. **In the official human evaluation, our system secured six first-place finishes, four second, and two third-place results** in the constrained category across the 13 directions we participated in. Notably, for the **English-Chinese**, our results surpassed all open/closed-source systems.

Lang. Pair	Human Evaluation
en→zh	Rank 1
en→ja	Rank 1
en→uk	Rank 1
ja→zh	Rank 1
en→bho	Rank 1
en→et	Rank 1
en→cs	Rank 2
en→ko	Rank 2
en→ru	Rank 2
cs→de	Rank 2
en→arz	Rank 3
cs→uk	Rank 3

Table 1: Human evaluation rankings of Marco-MT-Algharb at WMT2025.

1 Introduction

The Conference on Machine Translation (WMT) continues to be the primary arena for benchmarking the advancements in machine translation technology (Kocmi et al., 2024; Freitag et al., 2023). For years, the field was dominated by the Transformer architecture (Vaswani et al., 2017), which set a high standard for translation quality through its powerful attention mechanism. However, the recent advent of Large Language Models (LLMs) has sparked a paradigm shift. These models, pre-trained on vast amounts of text data, have demonstrated remarkable capabilities in understanding context, generating fluent text, and leveraging world knowledge, making them exceptionally promising candidates for complex multilingual translation tasks (Achiam et al., 2023; Ouyang et al., 2022; Ming et al., 2024; Alves et al., 2024). However, adapting these powerful, general-purpose LLMs for high-fidelity, specialized translation presents a significant challenge (Jiao et al., 2023; Hendy et al., 2023). To bridge this gap, we propose quality-aware training and decoding techniques designed to systematically enhance both the fluency and faithfulness of LLM-based translation. To this end, we present the Marco-MT-Algharb system.

Marco-MT-Algharb is our submission to the WMT 2025 General Machine Translation Shared Task. Our participation covers 13 diverse language pairs.[1] Built upon the Qwen3-14B foundation

[1]The 13 language pairs are: English to Chinese (en→zh), English to Japanese (en→ja), English to Korean (en→ko), English to Egyptian Arabic (en→arz), English to Bhojpuri (en→bho), English to Czech (en→cs), English to Estonian (en→et), English to Russian (en→ru), English to Ukrainian (en→uk), English to Serbian (Latin script) (en→sr_Latn), Czech to German (cs→de), Czech to Ukrainian (cs→uk), and Japanese to Chinese (ja→zh).

Proceedings of the Tenth Conference on Machine Translation, pages 587–593
November 8-9, 2025

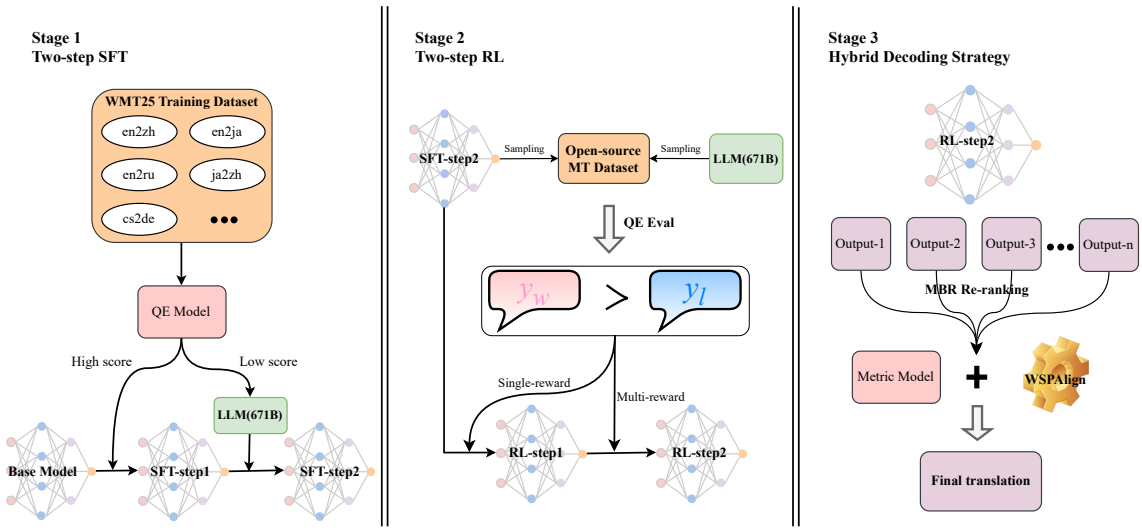


Figure 1: Overall pipeline of the Marco-MT-Algharb system. **SFT Stage:** Fine-tuning on QE-filtered and teacher-distilled data. **RL Stage:** Preference alignment via CPO and dynamic multi-reward optimization. **Decoding Stage:** Hybrid MBR re-ranking with a WSPAlign faithfulness score.

model (Yang et al., 2025), we propose quality-aware training and decoding techniques to enhance translation quality through three key phases: (1) a two-step Supervised Fine-Tuning (SFT) process with data distillation to expand data coverage; (2) a two-step Reinforcement Learning (RL) framework to align the model with quality estimation metrics; and (3) a hybrid decoding strategy that combines quality scores with word alignment to improve faithfulness.

Our work makes several key contributions to the development of state-of-the-art, LLM-based translation systems:

- **A progressive, two-step supervised fine-tuning (SFT) strategy.** We begin by training on high-quality, rigorously cleaned parallel data. Subsequently, we employ data distillation, using a powerful teacher model to regenerate translations for the data filtered out during the cleaning process. This allows our model to learn from a broader data distribution without being compromised by noise.
- **A two-step reinforcement learning (RL) framework for preference alignment.** We first utilize Contrastive Preference Optimization (CPO) (Xu et al., 2024) for initial alignment. We then introduce a novel dynamic multi-reward preference optimization method, which leverages a combination of quality metrics to achieve a more holistic improvement in translation adequacy and fluency.
- **A novel hybrid decoding strategy to mitigate omission errors.** We observed that models optimized heavily on neural quality estimation (QE) metrics like COMET (Rei et al., 2020) can sometimes produce fluent but incomplete translations. To address this, we developed a hybrid decoding algorithm that integrates a word-alignment-based penalty into the Minimum Bayes Risk (MBR) re-ranking framework, ensuring both semantic fidelity and lexical completeness.

In the official human evaluation (Kocmi et al., 2025), our system achieved top-3 rankings in 10 out of 13 participated directions within the constrained category, including five first-place finishes (see Table 1). Notably, our English-Chinese system surpassed all other submissions, including proprietary systems.

The remainder of this paper is structured as follows: Section 2 provides a detailed overview of our system's architecture and training methodology. Section 3 presents our experimental setup and results. Finally, Section 4 concludes the paper.

2 System Overview

Our translation system, Marco-MT-Algharb, is an end-to-end pipeline built upon a powerful foundation model. The overall workflow consists of four key stages: (1) selection of a base model architecture; (2) a two-step SFT process to impart translation capabilities; (3) a two-step RL process to align outputs with automated translation quality metrics; and (4) a hybrid decoding strategy for

robust and accurate inference. A schematic of our system is shown in Figure 1.

2.1 Model Architecture

We selected Qwen3-14B-base[2] as the foundation for our system. Qwen3 is a series of advanced, multilingual large language models known for their strong performance across a wide range of natural language understanding and generation tasks. The 14-billion parameter variant provides a powerful balance between model capacity and computational feasibility for fine-tuning. Its extensive pre-training on a diverse corpus of multilingual data makes it an excellent starting point for developing a high-quality, multilingual translation system, as it already possesses a rich cross-lingual representation space. For our tasks, we utilized the base model and tailored it specifically for translation through the subsequent training stages.

2.2 Two-step Supervised Fine-tuning

The goal of our SFT process is to effectively adapt the general capabilities of Qwen3-14B to the specific domain of machine translation. We designed a two-step approach to maximize data utilization and model performance.

Step 1: SFT on High-Quality Parallel Data. In the first step, we focused on building a robust translation baseline using high-quality data. We collected all parallel data provided by the WMT 2025 organizers for the 13 language directions we participated in. This raw data underwent an intensive cleaning pipeline, which included:

- Normalization: Standardizing punctuation, spacing, and casing.
- Filtering: Removing sentence pairs based on length ratio mismatches and identifying sentence pairs with a high proportion of non-alphabetic characters.
- Language Identification: Ensuring that the source and target sentences correctly match their designated language labels.
- Quality Estimation Filtering: Employing a pre-trained QE model (CometKiwi-XXL[3]) to score sentence pairs and discarding those with predicted low translation quality.

After cleaning, the resulting high-quality dataset was used to perform a single, comprehensive multilingual SFT run. For 12 of our high-resource language directions, we compiled a substantial dataset of 10 million parallel sentences each. Due to data scarcity for the English-to-Bhojpuri (en→bho) direction, its data volume was significantly smaller. In total, this initial SFT step utilized a massive training corpus of approximately 120 million sentence pairs, training the model on all 13 language pairs simultaneously. This large-scale multilingual training encourages the model to develop robust shared representations and effectively leverage cross-lingual transfer learning.

Step 2: SFT with Distilled Noisy Data. A significant amount of data is typically discarded during aggressive cleaning. While noisy, this data often contains valuable lexical and syntactic diversity. To leverage this, we designed a second SFT step based on data distillation. We took the parallel data that was filtered out in Step 1 and used a powerful teacher model, DeepSeek-V3[4] (Liu et al., 2024), to regenerate the target-side translations. For each of the 13 language directions, we distilled approximately 800,000 sentence pairs. This process effectively "cleans" the noisy target text while preserving the original source text's diversity. The resulting distilled dataset was then used for a second round of SFT. This allowed our model to learn from a much broader set of source inputs, guided by the high-quality outputs of the teacher model, thereby enhancing its robustness and domain coverage.

2.3 Preference Alignment via Two-step Reinforcement Learning

Following SFT, we employed a two-step RL to further refine the model's output. The goal of this stage is to directly align the model's generations with scores from automated translation quality estimation metrics, which serve as a proxy for human judgment. This approach moves beyond the token-level supervision of SFT to optimize the holistic quality of the entire translated sentence based on established evaluation standards.

Step 1: Contrastive Preference Optimization with Diverse Candidate Translations. We began with CPO to align our model with high-quality translation preferences. The foundation for our RL training is a curated dataset of source sentences, created by randomly sampling 15,000 entries for each source language from high-quality, open-source

[2]https://huggingface.co/Qwen/Qwen3-14B-Base

[3]https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

[4]https://huggingface.co/deepseek-ai/DeepSeek-V3

corpora, including Flores-200 (nll, 2024) and historical WMT test sets (WMT08-23). We reuse this same dataset for both of our RL steps, not only for methodological consistency, but more importantly, due to the limited availability of high-quality data that closely mirrors the test domain.

To construct preference pairs for these source sentences, we adopted a teacher-augmented strategy. For each source sentence, we generated a pool of candidate translations populated from two distinct sources: (1) multiple sampled outputs from our own SFT-trained model, and (2) a high-quality translation from a powerful, external teacher model, DeepSeek-V3.

We then used a reference-free QE model CometKiwi-XXL to score every candidate in this combined pool. The preference pair was formed as follows:

- The **"chosen"** translation was the candidate with the highest evaluation score. In many cases, this was the output from the DeepSeek-V3 model, providing a strong, high-quality target.
- The **"rejected"** translation was a candidate from the same pool with a significantly lower evaluation score, often one of the less successful samples from our own model.

This teacher-augmented approach is highly effective as it provides a robust learning signal, allowing our model to learn from responses that are often superior to its own initial capabilities. This CPO step provided a stable initial alignment towards the quality standards set by a strong translation model.

Step 2: Dynamic Multi-Reward Optimization with Self-Distillation. To achieve more nuanced control and move beyond reliance on a single, static metric, we introduce a novel training framework in our second RL step. This framework combines a dynamic, hybrid reward signal with a knowledge distillation objective, encouraging the model to internalize the principles of translation quality.

First, our reward function is not static but a dynamic composite of two sources: an external QE metric (CometKiwi-XXL) and the model's own internal reward signal. The total reward R_{total} for a generated translation y from source x at training step t is defined as:

$$R_{\text{total}}(x, y, t) = (1 - w_{\text{self}}(t)) \cdot R_{\text{QE}}(x, y) + w_{\text{self}}(t) \cdot R_{\text{self}}(x, y) \quad (1)$$

where R_{QE} is the score from the QE model, and $R_{\text{self}}(x, y)$ is the model's own sequence log-probability ($\log P_\theta(y|x)$), serving as a measure of its confidence. The weight of the self-reward, $w_{\text{self}}(t)$, is annealed to increase gradually with the training step t. This curriculum strategy allows the model to initially anchor its learning on the reliable external metric and progressively trust its own refined judgment as it improves.

Second, to accelerate the refinement of the model's internal judgment, we introduce a Kullback-Leibler (KL) divergence loss term. This term explicitly distills the relational quality knowledge from the QE model into the model's probability space. For a pair of translations (y_w, y_l) where QE scores y_w higher than y_l, we define a target preference distribution based on their score difference. The KL loss then minimizes the divergence between the model's predicted preference probability and this QE-derived target distribution:

$$\mathcal{L}_{\text{KL}} = D_{\text{KL}}\left(\sigma(\tau \cdot \Delta\text{QE}) \parallel \sigma(\Delta \log P_\theta)\right) \quad (2)$$

where ΔQE is the difference in QE scores for the pair (y_w, y_l), $\Delta \log P_\theta$ is the difference in the model's log-probabilities for the same pair, σ is the sigmoid function, and τ is a temperature parameter controlling the sharpness of the target distribution.

The final objective combines the preference optimization loss with $\mathcal{L}_{\text{KL}}$. This synergistic approach allows the model to not only generate better translations based on the hybrid reward but also to simultaneously internalize the very principles of translation quality evaluation. This makes the self-reward signal more reliable over time and leads to significant, stable performance gains.

2.4 Hybrid Decoding Strategy

A notable pitfall of optimizing Large Language Models towards neural metrics is their tendency to produce translations that are highly fluent yet lexically or semantically incomplete (Freitag et al., 2022; Moghe et al., 2022). To combat this, we developed a hybrid decoding strategy that fuses the MBR re-ranking algorithm with a reward for lexical faithfulness.

Our approach is built upon Minimum Bayes Risk (MBR) decoding (Freitag et al., 2021). In standard MBR, we first generate a set of N candidate translations $\{y_1, y_2, \ldots, y_N\}$ for a given source text x. The optimal translation y^* is the one that has the highest expected utility score against all other can-

Model	AVG	en→zh	en→arz	en→bho	en→cs	en→et	en→ja
Proprietary Models							
GPT-4o	74.20	75.70	74.85	30.19	85.44	76.63	78.35
Claude 3.7 Sonnet	74.24	76.79	76.08	29.12	85.10	76.71	79.48
Ablation Baselines (Marco-MT-Algharb)							
Two-step SFT	76.47	77.67	77.86	32.41	87.82	79.62	81.55
++Two-step RL	77.96	80.59	79.43	34.21	88.90	80.45	82.88
++Hybrid Decode	**79.33**	**82.39**	**80.13**	**38.61**	**90.50**	**83.39**	**83.24**

Model	en→ko	en→ru	en→uk	en→sr	cs→uk	cs→de	ja→zh
Proprietary Models							
GPT-4o	81.14	79.07	76.28	77.88	81.08	82.01	65.94
Claude 3.7 Sonnet	82.32	79.57	76.54	77.71	81.38	80.67	63.69
Ablation Baselines (Marco-MT-Algharb)							
Two-step SFT	84.52	81.35	78.49	80.58	82.47	82.81	66.99
++Two-step RL	84.50	82.87	80.02	82.02	83.37	83.54	68.27
++Hybrid Decode	**84.60**	**84.46**	**82.70**	**83.66**	**83.89**	**84.29**	**69.44**

Table 2: Main results on the WMT25 General test set, evaluated using XCOMET-XXL scores. We report the average (AVG) over all 13 language pairs. The best score in each column is in **bold**. For brevity, **en→sr** refers to the English-to-Serbian (Latin) direction.

didates:

$$y^* = \underset{y_i}{\text{argmax}} \sum_{j=1}^{N} U(y_i, y_j) \tag{3}$$

where the utility function $U(y_i, y_j)$ is realized using the COMET-22 metric model[5].

Our innovation is to incorporate an alignment-based score into this framework to explicitly reward source faithfulness. The final score for each candidate is a hybrid of its peer-based MBR utility and its source-based alignment score. We define the standard MBR score for a candidate y_i as:

$$S_{\text{MBR}}(y_i) = \sum_{j=1}^{N} U(y_i, y_j) \tag{4}$$

The alignment score, $S_{\text{align}}(x, y_i)$, is computed using our tool, `WSPAlign` (Wu et al., 2023), which measures the lexical faithfulness between the source x and the candidate y_i. A higher score indicates better faithfulness. The final hybrid score is then calculated as:

$$S_{\text{hybrid}}(y_i) = S_{\text{MBR}}(y_i) + \lambda \cdot S_{\text{align}}(x, y_i) \tag{5}$$

where λ is a hyperparameter that balances the MBR and alignment terms. During inference, we generate N candidates, compute S_{hybrid} for each, and select the one with the highest score as the final translation. This approach significantly reduces the frequency of omission errors in our final submissions.

3 Experiments

3.1 Experimental Setup

Implementation Details. For the Supervised Fine-Tuning (SFT) stage, we perform full-parameter fine-tuning. In contrast, for the Reinforcement Learning (RL) stage, we employ a parameter-efficient LoRA strategy (Hu et al., 2022), configuring the adapters with a rank of 64 and an alpha of 128. For the optimization process, we use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We adopt a carefully designed learning rate schedule that decreases with each stage of our pipeline: the learning rate was set to 2×10^{-5} for the first SFT step, 1×10^{-5} for the second SFT step, 2×10^{-6} for the first RL step (CPO), and 1×10^{-6} for the final RL step. A global batch size of 64 was maintained throughout all training phases. For our dynamic multi-reward

[5]We use the `Unbabel/wmt22-comet-da` implementation from Hugging Face.

optimization step, key hyperparameters include a KL distillation temperature τ of 0.3, and a self-reward weight w_{self} that is linearly annealed from an initial value of 0.05 to 0.8. In the subsequent Hybrid Decoding stage, the balancing weight λ is set to 0.5.

Evaluation Setup. We evaluate our final model on the official test sets for the WMT25 General Machine Translation Shared Task. Our participation covers 13 language pairs: en→{zh, arz, bho, cs, et, ja, ko, ru, uk, sr_Latn}, cs→{uk, de}, and ja→zh. Our decoding procedure implements the proposed hybrid re-ranking strategy. For each source sentence, we generate 100 candidate translations via stochastic sampling using the `vllm` library (Kwon et al., 2023) for efficient inference, which are then scored and re-ranked using our hybrid scoring function (Equation 5) to select the final output. The quality of this final translation is measured using the state-of-the-art reference-free metric XCOMET-XXL[6] (Guerreiro et al., 2024).

3.2 Main Results

Baselines. To demonstrate the effectiveness of our multi-stage pipeline, we compare our final system against several key baselines. We conduct an ablation study with two internal models: (1) Two-step SFT, our model after only the two SFT step, to measure the combined impact of our RL and hybrid decoding steps; and (2) Two-step RL, the model after full SFT and RL training, to isolate the performance gain from the hybrid decoding strategy. Furthermore, we benchmark our system against leading proprietary models, including GPT-4o and Claude 3.7 Sonnet.

Table 2 presents the main findings of our evaluation. The results clearly demonstrate the superiority of our final Marco-MT-Algharb system. On average, our final system achieves an XCOMET-XXL score of 79.33, significantly outperforming all other models in the comparison.

The effectiveness of our multi-stage pipeline is validated by the ablation study. Our full RL framework (Two-step RL) improves upon the SFT-only baseline by a substantial margin of 1.5 points on average. The final addition of our hybrid decoding strategy (Hybrid Decode) provides a further 1.3-point gain, highlighting the crucial and cumulative contribution of each component to the final performance.

[6]https://huggingface.co/Unbabel/XCOMET-XXL

Most notably, Marco-MT-Algharb not only surpasses the strong proprietary baselines of GPT-4o and Claude 3.7 Sonnet by a large margin (over 5.1 points on average), but it also achieves the highest score across every individual language pair. This consistently strong performance highlights the effectiveness of our specialized training and decoding approach. These results suggest that a carefully refined, open-source model can produce translations of exceptional quality, capable of outperforming even leading general-purpose proprietary systems on this benchmark.

4 Conclusion

This paper presented Marco-MT-Algharb, our system for the WMT25 General Machine Translation Shared Task. We introduced a novel quality-aware framework that enhances LLM-based translation through a synergistic combination of two-step supervised fine-tuning with data distillation, dynamic multi-reward reinforcement learning, and a hybrid alignment-aware decoding strategy. Our methodology was validated on the WMT25 test set, where Marco-MT-Algharb substantially outperformed strong baselines and leading proprietary models. These results were corroborated by the official human evaluation, which placed our system in the top three across 12 of our 13 language pairs, including six first-place victories. Notably, in the highly competitive English-Chinese direction, our system ranked first among all open and closed-source submissions.

References

2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021. Minimum bayes risk decoding with neural metrics of translation quality. *arXiv preprint arXiv:2111.09388*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, and 1 others. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, and 1 others. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *arXiv preprint arXiv:2412.04003*.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2022. Extrinsic evaluation of machine translation metrics. *arXiv preprint arXiv:2212.10297*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. Wspalign: Word alignment pre-training via large-scale weakly supervised span prediction. *arXiv preprint arXiv:2306.05644*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Evaluation of Qwen3 for English to Ukrainian Translation

Cristian Grozea[1] * Oleg Verbitsky[2]
[1]Fraunhofer Institute FOKUS, Berlin, Germany
[2]Pidstryhach Institute for Applied Problems of Mechanics and Mathematics
Ukrainian Academy of Sciences, Lviv, Ukraine

Abstract

We report the results of evaluating Qwen3 for the English-to-Ukrainian language pair of the general MT task of WMT 2025.

In addition to the quantitative evaluation, we performed a qualitative evaluation, in collaboration with a native Ukrainian speaker - therefore we present an example-heavy analysis of the typical failures the LLMs still do when translating natural language, particularly into Ukrainian.

We report also on the practicalities of using LLMs, such as on the difficulties of making them follow instructions, on ways to exploit the increased "smartness" of the reasoning models while simultaneously avoiding the reasoning part improperly interfering with the chain in which the LLM is just one element.

1 Introduction

This submission to the general MT task of WMT 2025 from Fraunhofer FOKUS continues our series of baselines prepared for the WMT biomedical translation task, as this system was intended for preparing baselines in case the task had been organized this year. Those baselines evolved with the field: starting with training sequence-to-sequence models on biomedical corpora (Bawden et al., 2019), continuing with transformers trained by us on biomedical corpora (Bawden et al., 2020), starting using generic pretrained models like T5 – not necessarily targeting the biomedical field (Yeganova et al., 2021; Neves et al., 2022), then using one of the first widely deployed LLMs, ChatGPT 3.5 (Neves et al., 2023) and eventually using the emerging more capable locally-hosted open-model LLMs, LLama 3.1 70B (Neves et al., 2024).

Using locally hosted models keeps data away from cloud providers, leveraging the competitiveness of the locally-hosted large LLMs such as the Qwen3 (Yang et al., 2025) we used here. This is particularly important to ensure unfiltered access to the potential hiding in these models, as online services may further censor or otherwise limit the cloud-served models.

2 System Description

2.1 Hardware

The hardware this system ran on is part of the Fraunhofer FOKUS GPU cluster. It is a server with 8 Nvidia H100 SXM GPUs, each with 80 GB Video RAM (VRAM), 3 TB RAM, 192 CPU cores (384 threads). The power consumption of this server in idle mode is about 1800 W. Only 7 of the 8 GPUs were used to perform the translation task, at an average of 140 W/GPU, adding approximately 1000 W to the power consumption during the experiment.

2.2 Software

The GPU cluster is managed with Kubernetes[1]. To run the selected LLM, we have used the ollama system[2], ran as a Helm[3] deployment and limited to 7 of the 8 GPUs.

The translation itself has been controlled with a Python script that used the ollama Python API for calling the LLM, once per each paragraph to translate.

Translating the texts took approximately 12 hours.

2.3 LLM

The LLM we have employed is Qwen3, as available in the ollama library under the tag qwen3:235b-a22b-q4_K_M. It is a mixture of experts model, with 235.1e9 parameters, context length 262144, embedding length 4096. It has been used with temperature 0.6 (randomness level in decoding),

*cristian.grozea@fokus.fraunhofer.de

[1]https://kubernetes.io/
[2]https://ollama.com/
[3]https://helm.sh/

Proceedings of the Tenth Conference on Machine Translation, pages 594–598
November 8-9, 2025

top_k=20, top_p=0.95 and repeat_penalty=1. This is a quantized model, with Q4_K_M quantization, leading in average to the use of 4.84 bits for every parameter of the original not-quantized model.

2.4 Prompt

This is the prompt we used uniformly, regardless of the context (be it chat, or video captioning or another source):

"You are a helpful assistant specialized in translation. You will be provided with a text in English, and your task is to translate it into Ukrainian. Keep the formatting as close as possible to the source. Preserve meaning, tone, emotions, and nuances and target the cultural context of Ukrainian. For easier selection, mark the translated text with <translation-begin> and <translation-end>".

3 Results and Analysis

3.1 Quantitative Evaluation

The automated evaluation performed by the organizers(Kocmi et al., 2025) computed several metrics characterizing the performance of our system: a trained quality estimation (CometKiwi-XL) score of 0.597 (versus 0.65 for the top system), an LLM evaluation score with the Commander-A model as judge of 75.7 (versus 84.1 for the top system), and of 78.1 with GPT4.1 as judge (versus 85.3 for the top system). A metric based on the reference translation (xCOMET-XL) was computed as well; score 0.513 (versus 0.662 for the top system). These metrics resulted in an AutoRank score of 8.7.

3.2 Qualitative Evaluation

We evaluated, grouped, and classified a randomly selected subset of errors on the very test set of the WMT 2025 general MT task, hoping that the recency of this dataset prevented the LLMs from being trained on these texts. We attempted to understand how the error came to be, given the statistical nature of the LLMs.

Types and sources of errors (**be warned, Example 3 could be shocking to sensitive readers, who might want to skip reading it**; the misspelled words in the source of translation are reproduced verbatim):

- The absence of idiomatic analogs in Ukrainian:

  Example 1

  EN: Rmember, theres alot of fish in the sea!

  UKR: Пам'ятай, у морі багато риби!

  This is a word-for-word translation. Ukrainians do not say it this way. However, in internet there are a lot of translations like this, which indicates this might become idiomatic in Ukrainian as well!

  Example 2

  EN: wejt out of their way

  UKR: вийшов зі свого шляху

  This is again an incorrect word-for-word translation. Correct would be: полишив свої справи.

- Unexplainable typos:

  Example 3

  EN: boil on my ass

  UKR: присом на сракі

  Correct: прищем на сраці

- (questionable) preservation of the English acronym:

  Example 4

  EN: WTF

  UKR: WTF (not translated)

- Play of words leading to the loss of sense:

  Example 5

  EN: Basically if you didn't know what that is,it's basically an example of two sentences. One says "Lets eat granny" making it seem like someone's gonna eat their nan.However,the other example says "Lets eat,granny" implying a different meaning to the sentence. Pretty creative way to teach kids to use commas ig.

  UKR: По суті, якщо ти не знав(-ла), що це таке, це просто приклад двох речень. Одне каже: «Їдьмо, бабуся», через що здається, ніби хтось збирається з'їсти свою бабусю. А інший приклад: «Їдьмо, бабусю», — що змінює зміст речення. Досить творчий спосіб навчити дітей ставити коми, припускаю.

  Mistake: "Їдьмо" is not "Let's eat" (though this incorrect guess is still somehow plausible/natural, due to the closeness of the stems of "to go" and "to eat" in Slavic languages). The actual meaning of "Їдьмо" is

"let's go/drive", but even then "to drive someone" does not exist in Ukrainian. Thus, the sense is completely lost. Correct: "Давайте їсти бабусю" and "Давайте їсти, бабусю". This word-for-word translation would perfectly preserve the general meaning of the original.

Example 6

EN: One word I simply can't say properly is water..... It actually hurts my brain trying to add the T and I sound american if I do or still say wut-ta, IDK why it's so hard for me to say tbh. Just wondering what ppl thought are when you meet people who drop H's and have glottal stops.

UKR: Одне слово, яке я просто не можу правильно вимовити – це вода... Навіть більно у голові, коли намагаюся додати Т, і виглядає, ніби я американець, якщо так роблю, або все ще кажу "вутта". Не знаю, чому мені так важко це вимовити, щиро кажучи. Цікаво, що люди думають, коли зустрічають тих, у кого випадає "г" і є гортанні зупинки.

The sense is lost as the word "вода" (translated from "water") does not contain the letter T at all.

- Translating with a blend of Ukrainian and Russian, including made-up words:

  Example 7

  EN: innit

  UKR: а?

  Correct: га? Or even better: чи не так? правда ж?

  Example 8

  EN: BLOW

  UKR: ЗІРВЕТЬСЯ

  Correct: ВИБУХНЕ.

  ЗІРВЕТЬСЯ here seems a blend of Ukrainian and Russian stemming from the Russian word ВЗОРВЁТСЯ. In fact, ЗІРВЕТЬСЯ exists in Ukrainian but has a completely different meaning (to fall off).

- corrupted words (in the translation)

  Example 9

  EN: RN

  UKR: ПРЯМОСЬОГО ТЕПЕР

  Correct: ПРЯМО ТЕПЕР.

  The word ПРЯМОСЬОГО doesn't exist in Ukrainian, it sounds slightly like ПРЯМІСІНЬКО (infrequent, more stringent variant of ПРЯМО).

  Yet another case appears in the translation from Example 6 above, "більно" does not exist in Ukrainian, although it is formed in a plausible way: біль+но. Correct would be боляче (painful).

- missing slang equivalents

  EN: BRUH

  UKR: БРУХ

  This is a transliteration, such a word does not exist in Ukrainian.

4 Discussion

When selecting Qwen3 235B, we evaluated it briefly and informally against another truly large LLM that can be run locally on powerful machines, DeepSeek-R1 671B (DeepSeek-AI, 2025), also quantized, with the same type of quantization. Somewhat surprisingly, the Qwen3 model, which is nearly three times smaller, seemed to outperform the largest DeepSeek-R1 model on English to Romanian and English to German tasks. In the introductory blog entry[4] Qwen3 has been presented as supporting 119 languages, including Ukrainian and English, whereas the training of DeepSeek-R1 focuses on English and Chinese (DeepSeek-AI, 2025). All this made us employ Qwen3 instead of DeepSeek-R1, as the more efficient Qwen3 was also faster. In retrospect, this seems to have been a poor decision, as the organizers report better results with DeepSeek-R1 for the task of English-to-Ukrainian translation. We should have contrasted the performance of those two models on the intended language pair.

Fine-tuning a general LLM to a task like this language pair translation was an option with the smaller models, but it became much more difficult as the models grew towards the limits of the available hardware resources, especially VRAM. The hope is not to have to specialize the models, but to get good results from generalist models instead. As the system we employed was intended as a baseline,

[4] https://qwenlm.github.io/blog/qwen3/

we refrained from attempting to improve upon the standard pre-trained Qwen3 model, as published for everyone.

Despite the instructions shown in Section 2.4, the LLM we employed did not always preserve the newline structure of the source, where double newlines served as paragraph delimiter. Therefore, and to make the task easier for the LLM, although with the risk of providing too little context for the translation, we split explicitly in the Python script the text into paragraphs, translated each paragraph and recomposed the resulting text by joining the translations with the expected delimiter. Even with this procedure, 50 out of 1251 texts had to be retranslated, because the LLM introduced spurious double newlines inside the translation of single paragraphs, disrupting the correspondence of the input and the output paragraphs.

Qwen3 is a so-called "reasoning" model, meaning that before outputting the desired translation it produces "reasoning" text describing its approach and its doubts about the task. This part is delimited by the tags <think> and </think>. We deleted this part of the output by removing everything up to and including the closing </think> tag.

The LLM inconsistently used the required tags to mark the translation and separate it from various other comments it produced in addition to it. It produced randomly such alternative closing tags for the translation section: </end-translation> and </begin-translation>. Our Python script was designed to detect and accept also these alternative closing tags.

In the end, two of the outputs still contained traces indicating that something went wrong with the automatic extraction of the translated text, that is they still contained the <think> tag. We decided not to fix those, in order to get a realistic evaluation of what to expect when using a "reasoning" LLM for translation.

We expect the need for explicit well-controlled postprocessing – as described here – to remain, as none of the models we interacted with have complete adherence to the instructions.

Concerning the errors those models still produce in machine translation, it might be that successfully finetuning on a well-curated parallel corpora might eliminate some of them, but probably not all.

5 Conclusion

We have evaluated the use of one of the largest local LLMs for automatic translation of English to Ukrainian. Beyond the automated quantitative analysis performed by the task organizers, we have performed a qualitative analysis of several translation errors observed, and explained also the engineering issues one has to deal with when relying on LLMs for machine translation, such as the incomplete adherence to instructions and the subsequent need for postprocessing, especially when using a "reasoning" LLM. We conclude that, although neural machine translation of occasionally challenging texts in natural language has advanced significantly, the LLMs, as the other neural models before them, continue to be characterized by two aspects: smooth and polished output, convincing thanks to the good form, paired with instances where meaning is sometimes missed – or even reversed. Still, when a human with knowledge of the target and of the source languages is proofreading the outcome, the translation process can be significantly accelerated using such a local LLM.

6 Thanks

We thank the anonymous reviewers, whose suggestions led to improving the quality of this article.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online. Association for Computational Linguistics.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, et al. 2025. Preliminary ranking of wmt25 general machine translation systems. *arXiv preprint arXiv:2508.14909*.

Mariana Neves, Cristian Grozea, Philippe Thomas, Roland Roller, Rachel Bawden, Aurélie Névéol, Steffen Castle, Vanessa Bonato, Giorgio Maria Di Nunzio, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, and Antonio Jimeno Yepes. 2024. Findings of the WMT 2024 biomedical translation shared task: Test sets on abstract level. In *Proceedings of the Ninth Conference on Machine Translation*, pages 124–138, Miami, Florida, USA. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 biomedical translation shared task: Evaluation of ChatGPT 3.5 as a comparison system. In *Proceedings of the Eighth Conference on Machine Translation*, pages 43–54, Singapore. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Amy Siu, Roland Roller, Philippe Thomas, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, Giorgio Maria Di Nunzio, Federica Vezzani, Christel Gerardin, Rachel Bawden, Darryl Johan Estrada, Salvador Lima-lopez, Eulalia Farre-maduel, Martin Krallinger, Cristian Grozea, and Aurelie Neveol. 2022. Findings of the WMT 2022 biomedical translation shared task: Monolingual clinical case reports. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 694–723, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.

SYSTRAN @ WMT 2025 General Translation Task

Dakun Zhang, Yara Khater, Ramzi Rahli, Anna Rebollo and **Josep Crego**
SYSTRAN by ChapsVision
5 rue Feydeau, 75002 Paris (France)
{dzhang,ykhater,rrahli,arebollo,jcrego}@chapsvision.com

Abstract

We present an English-to-Japanese translation system built upon the EuroLLM-9B (Martins et al., 2025) model. The training process involves two main stages: continue pretraining (CPT) and supervised fine-tuning (SFT). After both stages, we further tuned the model using a development set to optimize performance. For training data, we employed both basic filtering techniques and high-quality filtering strategies to ensure data cleanness. Additionally, we classify both the training data and development data into four different domains and we train and fine-tune with domain specific prompts during system training. Finally, we applied Minimum Bayes Risk (MBR) decoding and paragraph-level reranking for post-processing to enhance translation quality[1].

1 Introduction

Large language models (LLMs) are increasingly used in machine translation, taking advantage of their deep understanding of both source and target languages. Their training typically involves two main stages: continued pretraining (CPT) and supervised fine-tuning (SFT).

In the pretraining stage, the model is exposed to massive amounts of unlabeled text and learns by predicting the next token in a sequence. This allows the model to acquire a broad understanding of language structure, grammar, general world knowledge, and reasoning patterns. In the supervised fine-tuning stage, the model is trained on task-specific labeled datasets, where each input is paired with reference output. This targeted training enables the model to follow instructions more precisely and handle specialized tasks, such as question answering, summarization, classification, and translation, with greater accuracy.

For the WMT25 general translation task, we began with the pretrained LLM EuroLLM-9B (Martins et al., 2025) and performed additional training using bilingual corpora containing only English–Japanese sentence pairs. In this stage, we employed the two aforementioned training approaches: continued pretraining (CPT) and supervised fine-tuning (SFT), and trained separate systems using each method. Following the first stage, a reduced development dataset was employed to fine-tune (FT) the systems to the WMT translation tasks. The training architecture is shown in Figure 1 (left side).

Before generating translations, we segment the WMT25 test set into individual sentences using the newline character ("\n"), as our systems are trained to operate at the sentence level.

During inference, we apply Minimum Bayes Risk (MBR) decoding and reranking of diverse translation hypotheses produced by our two models. For each input sentence, we generate up to 300 translation candidates by combining outputs from both trained models with variations in decoding prompts (Table 6), greedy/nucleus decoding (5-best), and zero-shot/few-shot examples. A quality estimation step is then applied to these hypotheses, discarding the worst 50% for each input. From the remaining candidates, MBR decoding is used to select the most promising translation, following the approach of Rei et al. (2024). Finally, the translated sentences are concatenated back into paragraphs, which are reranked using CometKiwi[2] (Rei et al., 2022), with the top-ranked paragraph selected as the final system output. The inference post-process architecture is shown in Figure 1 (right side).

As a result, our submission is an ensemble of two open-weight, sentence-level, English-to-Japanese translation models with a combined total of 18B parameters. The following sections describe the

[1]We released the classification model and translation models: https://huggingface.co/Systran/collections

[2]Unbabel/wmt23-cometkiwi-da-xl. All CometKiwi scores in this work were computed using this model.

Proceedings of the Tenth Conference on Machine Translation, pages 599–606
November 8-9, 2025

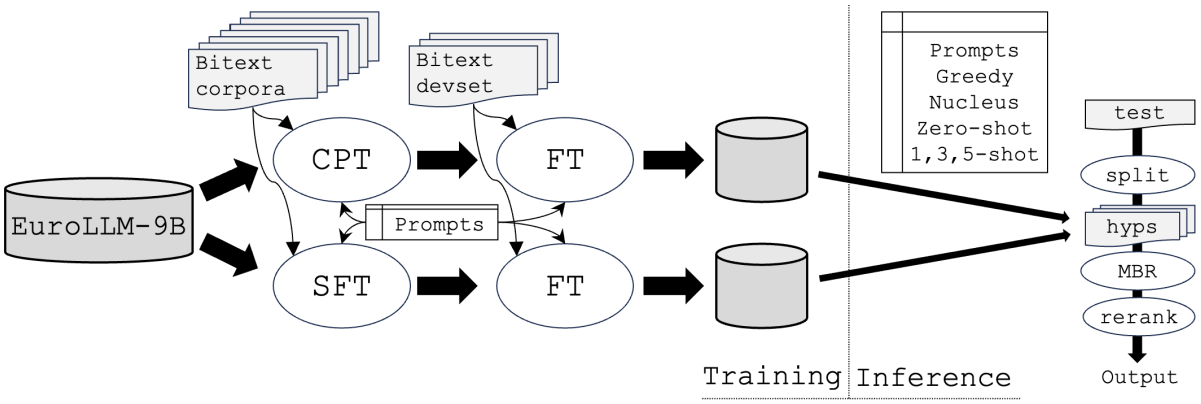


Figure 1: Frameworks for Training and Inference Post-Processing. System prompts are shown in Appendix A.

data preparation process as well as give additional details on the training procedures used to build our WMT25 English-Japanese system.

2 Data preparation

2.1 Bitext data

We use only parallel corpora to continue training LLMs for the machine translation task. We apply both basic and high-quality filtering methods to the WMT25 English-Japanese bitext data. From the original 225 million sentence pairs, we select approximately 10 million high-quality pairs for training. These filtered pairs are then used to train models using both CPT and SFT strategies. Similar filtering methods are also applied to the development datasets to further fine-tune (FT) the trained models. Table 1 summarizes the data statistics for the training and development datasets.

2.1.1 Basic Filtering Rules

Basic filtering rules are applied first to remove duplicates and noisy samples, such as misaligned sentence pairs and those with significantly different length ratios. Details of the filtering steps are as follows:

- **uniq:** Remove duplicated bitext examples.
- **1:1 example:** Discard examples where a single sentence is aligned to multiple sentences (1:N) or multiple sentences are aligned to a single sentence (M:1).
- **length ratio:** Discard examples where the length ratio between source and target exceeds 10 or is less than 0.2[3].
- **length:** Discard examples with length greater than 500 tokens.
- **character:** Retain sentences containing only Latin and Greek characters on the English side, and Latin, Greek, Hiragana, Katakana, and Han characters on the Japanese side.
- **LID:** Perform language identification using `lid.176.bin` (Joulin et al., 2016b,a).

2.1.2 High-Quality Filtering

The high-quality filtering process involves two main steps: first, generating translation hypotheses using a translation model; second, estimating the similarity between the source sentence and the target sentence, or between the source sentence and the generated hypothesis, using a quality evaluation model. For this purpose, we use EuroLLM-9B-Instruct (Martins et al., 2025) to translate English sentences into Japanese, and CometKiwi to assess the quality scores between English and Japanese segments in this year's WMT evaluation. Finally, we remove those samples which

$$CometKiwi(src, tgt) < CometKiwi(src, hyp)$$

2.2 Development data

Development data is used to fine-tune the pre-trained models to adapt to WMT evaluation. We first merge all development data, both English-to-Japanese and Japanese-to-English, provided by WMT25, including WMT dev/test data in previous years, NTREX (Federmann et al., 2022) and Flores (NLLB Team et al., 2024), except wmttest2024 dataset[4]. Then we apply high-quality filtering (Section 2.1.2) after deduplication. Details are shown

[3]We use basic tokenization for length and length ratio filtering: on the English side, we tokenize by spaces, while on the Japanese side, we tokenize at the character level.

[4]https://data.statmt.org/wmt24/general-mt/wmt24_GeneralMT.zip

	Bitext	Dev set
wmt25 provided	225M	20,111
uniq	132M	18,019
1:1 example	57M	-
basic rules (length, LID, etc.)	55M	-
high quality	10M	5,736

Table 1: Data statistics (number of lines) for training and development dataset filtering.

in Table 1. There are in total 5,736 sentence pairs finally used for system fine-tuning (FT in Table 1).

2.3 Data classification

In WMT2024, the test dataset covers four domains: News, Social, Speech and Literary. To perform domain classification, we fine-tune Llama-3.1-8B-Instruct[5] model on the WMT24++ dataset (Deutsch et al., 2025) and use the resulting classifier to label both the training and development data. The prompt used for classification during both training and inference is shown in Table 5.

The fine-tuned model categorizes English input sentences into one of the four domains. Within the 10M cleaned parallel corpus, the distribution across these domains is 7% (News), 74% (Social), 12% (Speech), and 7% (Literary), indicating that the Social domain constitutes the majority of the training data.

This classification model is used only for training data preparation. For decoding, we generate translation hypothesis for each input with all domain related prompts (Table 6) and rely on MBR postprocessing to select the best candidate.

Table 7 shows that domain related prompts benefit the final system.

3 Model training

3.1 Continue pretraining and Supervised fine-tuning

The pretraining stage of LLM typically requires vast amounts of unlabeled monolingual data. However, since the WMT evaluation is dedicated exclusively to machine translation, we leverage parallel corpora as a stand-in for monolingual data. Accordingly, we train LLMs independently using two approaches, continued pretraining (CPT) and supervised fine-tuning (SFT), on parallel data.

[5]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

For CPT, we generate training examples of up to 2048 tokens by appending the corresponding Japanese sentence to the end of each English input, together with a domain-specific prompt (Section 2.3). To minimize dependency between bilingual sentence pairs in the synthetic example, we insert an "end-of-sentence" token (</s>) after each Japanese sentence. The input text format for CPT is:

```
sample = (domain_prompt)En\nJa\n</s>
input = [sample] +
```

where `input` is constructed by concatenating (+) one or more `sample` entries, each containing an English (`En`) and a Japanese (`Ja`) sentence. The actual set of `domain_prompts` used are shown in Table 6.

We apply 10% prompt smoothing to the domain-specific prompts, where each of the four prompt types — News, Social, Speech and Literary — is sampled with a 10% probability with generic domain-free prompt (default 1 and default 2 in Table 6). This helps mitigate the impact of annotation errors and enhances training diversity.

We apply similar domain-specific prompts for SFT training. The only difference is that training examples are not concatenated during SFT.

We use LLaMA-Factory (Zheng et al., 2024) to train CPT/SFT models from EuroLLM-9B. The training parameters are the same as described in GemmaX2-28 (Cui et al., 2025) except that we use 4 GPUs in parallel for both training. The effective batch size for both trainings is 128. The learning rate starts from 2.0e-5 and decays based on cosine_with_min_lr policy, with a miminum value of 1.0e-6 (Table 8).

We use full model tuning rather than parameter-efficient methods such as LoRA adapters for CPT/SFT training. This choice is motivated by the relatively small size of the training data, where full tuning has been shown to yield better performance than adapter-based methods in similar low-resource settings (Hu et al., 2021; Pfeiffer et al., 2021).

3.2 Fine-tuning with development data

To further adapt our models to the WMT task, we perform an additional round of fine-tuning (FT) on both the CPT and SFT models using the filtered previous year's development data except wmttest2024 (Section 2.2).

Given the limited dataset size of 5,736 samples and an effective batch size of 128, this fine-tuning

Policy	Prompts	zero/few-shot	Greedy	Nucleus
Adequacy	all	all	Yes	No
Diversity	all	all	Yes	Yes

Table 2: Decoding policies (Adequacy vs. Diversity) for MBR post-processing.

wmttest2024		Adequacy policy		Diversity policy	
		BLEU	CometKiwi	BLEU	CometKiwi
Baseline	EuroLLM-9B-instruct	26.3	0.7501	22.9	0.7620
	EuroLLM-9B-instruct-FT	26.4	0.7500	22.6	0.7624
Our models	EuroLLM-9B-CPT-FT	29.5	0.7504	24.5	0.7659
	EuroLLM-9B-SFT-FT	28.4	0.7515	24.9	0.7667
	Ensemble	**29.8**	0.7586	24.6	**0.7686**

Table 3: Sentence-level evaluations on wmttest2024 English-Japanese translation. Both BLEU and CometKiwi scores are the result of MBR output. BLEU is calculated by Sacrebleu (Post, 2018) with ja-mecab tokenization. CometKiwi is the reference-free score of Unbabel/wmt23-cometkiwi-da-xl.

process runs for only 45 steps (1 epoch). In this stage, training samples are also built with domain-specific prompts as in our previous training. We apply full model training instead of using LoRa adapters for fine-tuning, with a starting learning rate 2.0e-5 and inverse_sqrt decay policy. We add NEFTune noise (Jain et al., 2023), with alpha=5, in this fine-tuning stage (Table 8).

4 Decoding and Postprocessing

We follow the idea of quality-aware decoding proposed by Rei et al. (2024). First, we apply Quality Estimation (QE) to discard the translation candidates with the lowest quality. Then, we perform Minimum Bayes Risk (MBR) decoding over the remaining hypotheses, using CometKiwi (Rei et al., 2022) as the loss function. The candidate with the lowest expected risk is selected as the final output.

To generate diverse output for each input, we apply different domain-free/domain-specific prompts, zero/few-shot examples, and greedy decoding as well as nucleus decoding. Table 2 summarizes the two decoding policies that we used in this work. Note that the policies differ only in the use of nucleus sampling to generate diverse hypotheses, which favors outputs with higher diversity (lower adequacy). Table 3[6] confirms this, as the Adequacy policy achieves correspondingly higher BLEU scores, while the Diversity policy yields higher CometKiwi scores.

[6]EuroLLM-9B-instruct-FT is the model that we directly fine-tune EuroLLM-9B-instruct with the development data described in Section 2.2.

4.1 Reference-free Quality Estimation (QE)

To filter low-quality hypotheses, we employ Comet-QE (Rei et al., 2021) in a reference-free setting. Given a source sentence src and a set of N candidate translations $\{hyp_1, hyp_2, \ldots, hyp_N\}$, we use Comet-QE score to compute quality scores $s_i = \text{Comet-QE}(src, hyp_i)$ for each hypothesis hyp_i. We then retain only the top $K = \lfloor N/2 \rfloor$ candidates with the highest scores:

$$\{hyp'_1, \ldots, hyp'_K\} = \text{Top-}K\ (\{s_1, \ldots, s_N\})$$

This step serves to remove noisy or low-quality translations that may adversely affect subsequent MBR decoding (Kondo et al., 2024).

4.2 Minimum Bayes Risk (MBR) Decoding

Following QE filtering, we apply Minimum Bayes Risk decoding (Fernandes et al., 2022) using a reference-based CometKiwi score. For each sentence, we consider the remaining K candidates $\{hyp_1, \ldots, hyp_K\}$ and compute the pairwise loss between each one of them with the others using CometKiwi scores, treating each candidate as a reference for the others: $\ell(hyp_i, hyp_j) = 1 - \text{CometKiwi}(src, hyp_i, hyp_j)$. Each hypothesis is then assigned an expected loss:

$$\mathbb{E}[\ell(hyp_i)] = \sum_{j=1}^{K} p(hyp_j) \cdot \ell(hyp_i, hyp_j)$$

where $p(hyp_j) = \frac{\exp(\log P(hyp_j))}{\sum_k \exp(\log P(hyp_k))}$ is derived from the log-probabilities assigned by the model.

wmttest2025	CometKiwi	
	Adequacy policy	Diversity policy
EuroLLM-9B-CPT-FT	0.6769	0.6801
EuroLLM-9B-SFT-FT	0.6826	0.6842
Ensemble	0.6843	0.6859
Reranking (main submission)	**0.7033**	

Table 4: Paragraph-level evaluation (Kocmi et al., 2025) for wmttest2025. CometKiwi is the reference-free score of Unbabel/wmt23-cometkiwi-da-xl.

We select the hypothesis y^* that minimizes the expected loss:

$$y^* = \arg\min_{hyp_i} \mathbb{E}[\ell(hyp_i)]$$

and thus, selecting the hypothesis that is most representative of the overall candidate distribution.

4.3 Paragraph-level re-ranking (WMT25)

For the WMT25 test set, we first split each document into sentences using the newline character "\n". Then we apply MBR to select the best candidate for each sentence individually. The selected sentences are reassembled into their original paragraph structure. Finally, we compute paragraph-level CometKiwi scores and select the combination of sentences that yields the highest overall score as the final output.

Let $D = \{d_1, d_2, \ldots, d_N\}$ be the set of documents in the WMT25 test set. Each document d is split into a list of sentences:

$$S_d = [s_1, s_2, \ldots, s_{n_d}]$$

For each sentence s_i, let $y_i = \{y_{i_1}, y_{i_2}, \ldots, y_{i_k}\}$ be a set of corresponding candidate translations. Apply MBR decoding to select the best candidate $\hat{y}_i$, and reconstruct the sentence list with best translation candidates:

$$\hat{S}_d = [\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{n_d}]$$

These selected sentences are then reassembled into a paragraph structure:

$$\hat{P}_d = \text{Assemble}(\hat{S}_d)$$

Finally, among all possible sentence combinations $P \in \mathcal{P}_d$, select the one with the highest paragraph-level CometKiwi score:

$$\hat{P}_d^* = \arg\max_{P \in \mathcal{P}_d} \text{CometKiwi}(P)$$

As shown in Table 4, incorporating paragraph-level reranking further improves the quality of the final submission.

5 Conclusions

In this paper, we present our English-to-Japanese translation system for the WMT25 General Translation Task. The final output is generated by ensembling two models trained with different strategies: continued pretraining (CPT) and supervised fine-tuning (SFT). Both models are trained on a cleaned parallel corpora of 10 million sentence pairs and further fine-tuned (FT) on a development set consisting of 5,736 sentences. MBR and re-ranking inference post-processing are also successfully performed to obtain the final quality boost. Additionally, we release one classification model[7] and two translation models[8] in our HuggingFace repository.

Acknowledgments

This research was funded by the French Agence Nationale de la Recherche (ANR) under the project TraLaLaM ("ANR-23-IAS1-0006"). This work was performed using HPC resources from GENCI–IDRIS (Grant 2024-A0161015117).

References

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages Dialects.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First*

[7] https://huggingface.co/Systran/Llama-3.1-8B-Instruct-ft-wmt25-classifier

[8] https://huggingface.co/Systran/EuroLLM-9B-cpt-ft-wmt25-en-ja, https://huggingface.co/Systran/EuroLLM-9B-sft-ft-wmt25-en-ja

Workshop on Scaling Up Multilingual Evaluation, pages 21–24, Online. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Neftune: Noisy embeddings improve instruction finetuning.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Preliminary ranking of wmt25 general machine translation systems.

Minato Kondo, Ryo Fukuda, Xiaotian Wang, Katsuki Chousa, Masato Nishimura, Kosei Buma, Takatomo Kano, and Takehito Utsuro. 2024. NTTSU at WMT2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 270–279, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Prompts

classifier	{"role": "system", "content": "You are a language expert."} {"role": "user", "content": "Classify the following English sentence into one of the following categories based on its content and style:\nNews: Factual reporting or informative text typically found in journalism.\nSocial: Informal, conversational, or casual text often used on social media or in personal messages.\nSpeech: Spoken or scripted verbal communication, such as political speeches, interviews, or lectures.\nLiterary: Creative or artistic writing, including fiction, poetry, or other literary works.\n\nSentence: {line}\n\nYour task: Identify the most appropriate category from the four above. Just respond with the category name: News, Social, Speech, or Literary."}

Table 5: Prompt used to fine-tune classification model and inference

default.1	Please translate the following {source_language} text into {target_language}.\nInput: {line}\nOutput:
default.2	You're a professional {target_language} translator. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.news	You're a professional {target_language} translator. You are translating a news article. The translation should be clear, objective, and concise, preserving factual accuracy and the original journalistic tone. Do not add opinions or interpretations. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.literary	You're a professional {target_language} translator. You are translating a literary text. Pay attention to stylistic features such as imagery, rhythm, and narrative voice. The translation should be faithful to the original tone and emotional depth, while adapting gracefully into the target language. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.speech	You're a professional {target_language} translator. You are translating a speech transcript. Maintain a fluent, persuasive tone suitable for public speaking. The output should be easy to read aloud and emotionally engaging, while staying faithful to the speaker's intent. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:
domain.social	You're a professional {target_language} translator. You are translating a social media post or informal message. The tone should be natural, conversational, and culturally relevant. Preserve emojis, slang, and informal expressions when appropriate. Please translate the following {source_language} sentence into {target_language}. You must only answer with the translation.\n{source_language} sentence: {line}\n{target_language} sentence:

Table 6: Prompt used to train translation model and inference

B Results on domain related prompts

EuroLLM-9B-CPT	+FT	wmttest2024	
		BLEU	CometKiwi
✗		20.7	0.7120
✓		22.9	0.7140
✗	✗	27.7	0.7373
✗	✓	28.1	0.7369
✓	✗	28.7	0.7377
✓	✓	28.6	0.7376

Table 7: Evaluation of the EuroLLM-9B-CPT-FT model trained with domain-related prompts (✓) and without domain-related prompts (✗). Scores are computed on wmttest2024 using the default prompt during the decoding phase.

C Training parameters

	CPT	SFT/FT
per_device_train_batch_size	4	4
Number of GPUs	4	4
gradient_accumulation_steps	8	8
Data cutoff length	2048	2048
Number of epochs	1	1
Max Learning Rate	2.0e-5	2.0e-5
Min Learning Rate	1.0e-6	/
lr_scheduler_type	cosine_with_min_lr	inverse_sqrt
Finetuning Type	full	full
bf16	true	true
Template	empty	empty
warmup_ratio	0.01	0.01
weight_decay	0.01	0.01
Optimizer	AdamW	AdamW
Deepspeed	ZeRO2	ZeRO2
neftune_noise_alpha	/	5

Table 8: The training parameters for CPT, SFT and FT are configured according to Cui et al. (2025).

Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, Mingyang Song

Tencent Hunyuan
moonzheng@tencent.com
https://github.com/Tencent-Hunyuan/Hunyuan-MT

Abstract

In this paper, we present our submission to the WMT25 shared task on machine translation, for which we propose **S**ynergy-en**h**anced polic**y** optimization framework, named **Shy**. This novel two-phase training framework synergistically combines knowledge distillation and fusion via reinforcement learning. In the first phase, we introduce a multi-stage training framework that harnesses the complementary strengths of multiple state-of-the-art large language models to generate diverse, high-quality translation candidates. These candidates serve as pseudo-references to guide the supervised fine-tuning of our model, Hunyuan-7B, effectively distilling the collective knowledge of multiple expert systems into a single efficient model. In the second phase, we further refine the distilled model through Group Relative Policy Optimization, a reinforcement learning technique that employs a composite reward function. By calculating reward from multiple perspectives, our model ensures better alignment with human preferences and evaluation metrics. Extensive experiments across multiple language pairs demonstrate that our model **Shy-hunyuan-MT** yields substantial improvements in translation quality compared to baselines. Notably, our framework achieves competitive performance comparable to that of state-of-the-art systems while maintaining computational efficiency through knowledge distillation and fusion.

1 Introduction

The field of machine translation has witnessed remarkable progress with the emergence of large language models; yet, challenges remain in consistently producing human-quality translations. This work addresses two key limitations: (1) the over-reliance on single-model supervision during fine-tuning, and (2) the difficulty of aligning machine outputs with nuanced human judgments during reinforcement learning.

Our method comprises three main phases. First, we collect diverse translations from state-of-the-art open-source large language models, including DeepSeek-V3 (DeepSeek-AI, 2024), DeepSeek-R1 (Guo et al., 2025), and Gemma across multiple language pairs. This ensemble approach provides a richer training signal than single-model distillation, exposing our base model Hunyuan-7B to varied translation strategies and stylistic choices. The collected translation outputs are carefully filtered and normalized before serving as supervision targets. Second, we perform Supervised Fine-Tuning (SFT) on Hunyuan-7B using the prior collected translation dataset. Crucially, we implement a dynamic weighting scheme that prioritizes higher-quality translations during SFT training, as determined by automatic metrics. Specifically, this phase enables the model to internalize the strengths of each contributor model while maintaining its own linguistic identity. The third phase applies Group Relative Policy Optimization (Shao et al., 2024), a sample-efficient Reinforcement Learning (RL) algorithm, to further refine the model. We employ XCOMET for its strong correlation with human judgments and DeepSeek-V3 for its complementary strengths in fluency assessment. The reward function combines these signals with a KL-divergence term to prevent excessive deviation from the SFT model. During RL training, we maintain multiple policy groups that explore different translation strategies, with periodic selection pressure favoring the approaches that yield the highest rewards.

Our methodology incorporates several primary techniques to ensure robust performance. Specifically, we employ a temperature-scaled sampling strategy during preference data collection to optimize the trade-off between diversity and quality. For the SFT phase, a layer-wise learning rate decay is applied to enhance model adaptation. During the RL phase, we implement a dynamic reward normalization scheme to maintain training stability.

Proceedings of the Tenth Conference on Machine Translation, pages 607–613
November 8-9, 2025

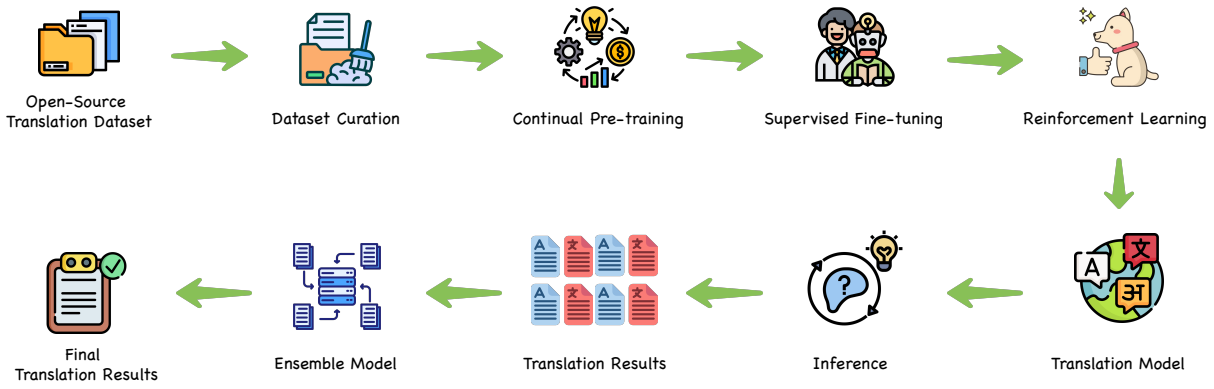


Figure 1: **Framework of the Shy-hunyuan-MT**. Firstly, we leverage an open-source translation dataset to conduct continuous pre-training on the Hunyuan-7B model. Secondly, we utilize the collected WMT translation data to perform SFT on the continuously pre-trained model. Thirdly, we sample outputs from the translation model to generate a set of diverse translation results, which are then used to train and derive the final translation result.

Model performance is rigorously evaluated on the WMT25 test sets using both automatic metrics and human assessment. Comprehensive ablation studies are conducted to validate the efficacy of each proposed component. The primary contributions are threefold:

- A novel and practical framework for the refinement of machine translation models.
- Compelling empirical evidence demonstrating the effectiveness of RL in the domain of machine translation.
- Valuable insights into the application and dynamics of multi-reward RL for complex text generation tasks.

2 Related Work

This section reviews the key research strands underlying our approach: (1) large language models for machine translation, (2) continual pre-training strategies, (3) supervised fine-tuning for translation tasks, and (4) reinforcement learning optimization for text generation.

2.1 Large Language Models for MT

The success of dense decoder-only models, such as GPT-3 (Brown et al., 2020), has revolutionized neural machine translation (NMT). Recent work has demonstrated that 7B-parameter models, such as Hunyuan, can achieve competitive performance when properly adapted. Unlike traditional encoder-decoder architectures (Vaswani et al., 2017), monolithic LLMs process translation as conditional text generation, offering advantages in zero-shot capability and multilingual transfer (Kirstain et al., 2021). Our work extends this direction by systematically investigating continual pre-training strategies for domain adaptation in translation tasks.

2.2 Continual Pre-training for MT

Domain adaptation through continual pre-training has shown promise in recent MT research. Gururangan et al. (2020) established that targeted pre-training on in-domain corpora improves downstream task performance. For WMT competitions specifically, Liao et al. (2021) demonstrates the effectiveness of iterative pre-training on parallel corpora. Our approach differs in that it employs a two-phase adaptation: first, on general bilingual corpora, and then on WMT historical data. This hierarchical adaptation strategy aligns with findings from (Pfeiffer et al., 2020) about the importance of gradual domain specialization.

2.3 Supervised Fine-tuning for MT

The transition from pre-training to task-specific fine-tuning remains a research area of active interest. Raffel et al. (2020) demonstrates that controlled fine-tuning with progressive data augmentation can prevent catastrophic forgetting. Our SFT protocol incorporates three key innovations: (1) dynamic batch sampling based on sentence complexity metrics, (2) gradient accumulation for low-frequency language pairs, and (3) temperature-annealed decoding during training. These techniques build upon curriculum learning principles first proposed by Bengio et al. (2009), but with specific adaptations for translation tasks.

2.4 Reinforcement Learning for MT

Recent advances, such as GRPO (Shao et al., 2024), provide more stable RL optimization for MT and various tasks (Guo et al., 2025; Song et al., 2025; Yang et al., 2025; Li et al., 2025; Zheng et al., 2025). To obtain better rewards, we use a combination of different metrics as the reward, and this multi-metric approach addresses limitations identified by Freitag et al. (2022) regarding single-metric optimization. The ensemble strategy further builds on the diversity-promoting techniques from Vijayakumar et al. (2016), but with novel modifications for temperature-controlled output variation.

2.5 WMT Competition Innovations

Analysis of prior WMT winning systems reveals evolving trends. The 2021 Edinburgh system (Chen et al., 2021) pioneered the use of large-scale back-translation, while Guu et al. (2020) demonstrates the effectiveness of retrieval-augmented models. Our work contributes to this lineage by showing how to effectively combine continual pre-training, reinforced fine-tuning, and learned ensemble strategies within a single LLM framework, addressing the scalability challenges noted by Tom et al. (2023) in their WMT 2023 overview.

3 Methodology

Our approach consists of two major phases: (1) a three-stage training pipeline for developing a high-quality base translation model, and (2) an ensemble strategy that leverages diversity generation and reinforcement learning to produce final translations. The overall architecture is illustrated in Figure 1.

3.1 Continual Pre-training

We leverage Hunyuan-7B, a state-of-the-art multilingual dense foundation model, as our initialization checkpoint. To effectively adapt this general-purpose model for machine translation tasks, we perform domain-adaptive continual pre-training using diverse large-scale parallel and monolingual corpora. Our training corpus encompasses multiple data sources with complementary characteristics:

- **OPUS Collection** (Tiedemann, 2012): A comprehensive multilingual parallel corpus covering over 20 language pairs across diverse domains, providing broad linguistic coverage
- **ParaCrawl** (Buck and Koehn, 2016): Large-scale web-crawled parallel data offering extensive real-world language usage patterns
- **UN Parallel Corpus** (Ziemski et al., 2016): High-quality formal documents ensuring exposure to professional and diplomatic language registers
- **C4** (Raffel et al., 2020): Cleaned English text derived from Common Crawl, contributing to robust monolingual understanding
- **WikiText** (Merity et al., 2016): Encyclopedic articles providing well-structured, factually accurate content

This diverse mixture enables our model to acquire comprehensive translation capabilities across various domains, registers, and language pairs, while maintaining the strong multilingual representations inherited from the base model.

3.2 Supervised Fine-tuning

Following domain adaptation, we conduct supervised fine-tuning using WMT benchmark datasets spanning from 2015 to 2024. Our training method incorporates several regularization techniques to mitigate catastrophic forgetting while preserving the model's acquired capabilities. Then, we optimize the standard sequence-to-sequence cross-entropy objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log p(y_t | y_{<t}, x; \theta) \tag{1}$$

where x denotes the source sentence, y represents the target translation, and θ encompasses the model parameters. Our training configuration employs the following strategies:

- **Learning rate scheduling**: We implement linear warmup over 5% of total training steps to ensure stable optimization dynamics.
- **Gradient regularization**: We apply gradient clipping with a maximum norm of 1.0 to prevent gradient explosion.
- **Computational efficiency**: We utilize mixed-precision training with BF16 representation to accelerate training while maintaining numerical stability.

This fine-tuning protocol enables effective task-specific adaptation while maintaining the robustness gained through domain pre-training.

3.3 Reinforcement Learning

After SFT, we apply GRPO, a sample-efficient reinforcement learning algorithm. We design the reward function by combining three metrics:

$$r = w_1 \cdot \text{BLEU} + w_2 \cdot \text{XCOMET} + w_3 \cdot \text{DeepSeek} \quad (2)$$

where $w_1 = 0.2$, $w_2 = 0.4$, and $w_3 = 0.4$ are empirically determined weights. Then we optimize the GRPO objective as follows:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \\ \frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i,\right.\right. \\ \left.\left.\text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \\ \beta\mathbb{D}_{\text{KL}}\left(\pi_\theta \parallel \pi_{\text{ref}}\right), \end{aligned} \quad (3)$$

$$\mathbb{D}_{\text{KL}}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \quad (4)$$

where ϵ and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}. \quad (5)$$

where A_i is the advantage function estimated using group-normalized rewards. To improve the translation quality of terminology and low-resource languages, we also use the methods proposed in TAT-R1 (Li et al., 2025), SSR-Zero (Yang et al., 2025), and Hunyuan-MT-7B (Zheng et al., 2025).

3.4 Synergy-based Policy Optimization

To further improve translation quality, we develop a two-phase fusion approach.

3.4.1 Diversity Generation

For each input sentence x, we generate $N = 5$ candidate translations by varying:

- Temperature ($\tau \in \{0.3, 0.5, 1.0, 1.5, 2.0\}$)
- Random seeds (5 different initializations)
- Beam search width (4-6 beams)

These settings allow our model to produce a candidate pool $Y = \{y_1, ..., y_N\}$ covering different translation properties (e.g., fluency vs. adequacy).

3.4.2 RL Training

We train another model to select/combine candidates from Y via GRPO. The framework utilizes the translation model as the base model in the first stage. The reward function remains the same weighted combination, now applied to the final output. During inference, the model can either:

1. Select the highest-scoring candidate;
2. Generate a new translation by attending to all candidates;

3.5 Implementation Details

All models are implemented in PyTorch and trained on 128 GPUs. Specifically, key hyperparameters are shown in Table 1.

Table 1: Training Hyperparameters

Parameter	Value
Batch size	64
Learning rate	5×10^{-5}
Max sequence length	8196

4 Results

Table 2 presents comprehensive evaluation results of our proposed Shy-hunyuna-MT model across 31 language directions from the WMT25 benchmark. The results of our model demonstrate exceptional performance across diverse linguistic pairs, achieving consistent state-of-the-art results on multiple automatic evaluation metrics. Our model achieves a remarkable AutoRank score of 1.0 across all 31 translation directions, indicating superior performance compared to baseline systems. This consistency across diverse language pairs, ranging from high-resource languages (e.g., English-German, English-Japanese) to low-resource ones (e.g., English-Bhojpuri, English-Maasai), demonstrates the robustness and generalization capability of Shy-hunyuna-MT.

Meanwhile, the evaluation results encompass both reference-based and reference-free metrics. On the CometKiwi-XL, our model achieves scores ranging from 0.577 to 0.720, with robust performance on the English-Estonian (0.720) and English-Korean (0.697) pairs. For GEMBA-ESA, our model consistently demonstrates excellence, with GEMBA-ESA-GPT4.1 scores predominantly

Table 2: Results of Shy-hunyuna-MT on 31 language directions of WMT25.

Direction	AutoRank↓	CometKiwi-XL↑	GEMBA-ESA-CMDA↑	GEMBA-ESA-GPT4.1↑	MetricX-24-Hybrid-XL↑	XCOMET-XL↑	chrF++↑
English-Egyptian Arabic	1.0	0.658	76.3	75.0	-5.7	0.388	-
English-Bhojpuri	11.5	-	-	-	-	-	40.6
English-Czech	1.0	0.658	83.7	89.4	-5.5	0.639	-
English-Estonian	1.0	0.72	78.8	87.8	-7.3	0.628	-
English-Icelandic	1.0	0.663	71.6	83.9	-7.5	0.543	-
English-Italian	1.0	-	84.6	88.7	-4.7	0.62	-
English-Japanese	1.0	0.687	82.2	89.6	-5.5	0.592	-
English-Korean	1.0	0.697	83.8	85.6	-4.9	0.624	-
English-Maasai	1.0	-	-	-	-	-	27.7
English-Russian	1.0	0.657	84.3	85.9	-4.9	0.652	-
English-Serbian (Cyrilics)	1.0	0.687	76.6	83.3	-4.2	0.64	-
English-Ukrainian	1.0	0.65	84.1	85.3	-5.0	0.662	-
English-Simplified Chinese	1.0	0.67	87.2	88.3	-4.0	0.576	-
Czech-Ukrainian	1.0	0.601	79.1	85.3	-5.0	0.681	-
Czech-German	1.0	0.596	78.4	88.3	-3.6	0.653	-
Japanese-Simplified Chinese	1.0	0.577	85.1	85.5	-4.2	0.629	-
English-Bengali	1.0	-	67.9	83.2	-4.8	0.449	-
English-German	1.0	-	84.3	90.6	-3.1	0.703	-
English-Greek	1.0	-	80.3	85.8	-5.3	0.601	-
English-Persian	1.0	-	80.4	84.1	-4.6	0.553	-
English-Hindi	1.0	-	77.0	82.3	-5.1	0.44	-
English-Indonesian	1.0	-	83.2	87.1	-4.4	0.677	-
English-Kannada	1.0	-	64.0	78.8	-6.0	0.446	-
English-Lithuanian	1.0	-	77.6	84.1	-6.3	0.569	-
English-Marathi	1.0	-	70.8	81.6	-5.8	0.248	-
English-Romanian	1.0	-	83.2	86.3	-5.7	0.651	-
English-Thai	1.0	-	71.3	87.9	-5.1	0.603	-
English-Serbian (Latin)	1.0	-	80.1	84.2	-3.4	0.583	-
English-Swedish	1.0	-	84.2	91.0	-4.7	0.685	-
English-Turkish	1.0	-	81.4	85.2	-7.2	0.542	-
English-Vietnamese	1.0	-	83.1	87.3	-4.5	0.623	-

above 80%, reaching peaks of 91.0% for English-Swedish and 90.6% for English-German, indicating high-quality translations that align well with human preferences. The MetricX-24-Hybrid-XL scores, while negative across all directions, remain relatively compact within the range of -3.1 to -7.5, with English-German achieving the best score of -3.1. This metric consistency indicates stable translation quality without significant degradation across different language families.

Furthermore, the proposed model Shy-hunyuna-MT exhibits powerful performance on European language pairs, with XCOMET-XL scores exceeding 0.65 for English-German (0.703), English-Swedish (0.685), and Czech-Ukrainian (0.681). For Asian languages, the results maintain competitive performance with Japanese-Simplified Chinese achieving 0.629 and English-Korean reaching 0.624 on XCOMET-XL. Notably, for low-resource languages such as English-Maasai and English-Bhojpuri, where most neural metrics are unavailable, the model still achieves chrF++ scores of 27.7 and 40.6, respectively, demonstrating its capability to handle challenging low-resource scenarios where traditional evaluation metrics fail to provide coverage. In addition, the model demonstrates effective cross-lingual transfer capabilities, as evidenced by its strong performance on non-English-centric pairs, such as Czech-Ukrainian, Czech-German, and Japanese-Simplified Chinese. These directions achieve comparable scores to English-centric pairs, with Czech-Ukrainian no-

tably achieving 0.681 on XCOMET-XL, surpassing many English-centric directions.

In summary, Shy-hunyuna-MT establishes new benchmarks across diverse translation directions, demonstrating both breadth in language coverage and depth in translation quality, making it a versatile solution for MT tasks.

5 Conclusion

In this work, we propose **Shy-hunyuan-MT**, a novel MT system built upon the **S**ynergy-en**h**anced polic**y** optimization framework (**Shy**). Our method leverages a carefully designed two-phase training paradigm that systematically transforms the open-sourced Hunyuan-7B base model into a state-of-the-art translation system. The core innovation of our method lies in the synergistic combination of three complementary training phases: domain-adaptive continual pre-training on large-scale parallel corpora, supervised fine-tuning on curated WMT datasets, and reinforcement learning through Generalized Reward Policy Optimization (GRPO) with composite reward signals. This progressive training strategy enables the model to acquire robust multilingual translation capabilities while maintaining strong performance across a diverse range of language pairs.

Our extensive experiments on 31 language directions from WMT25 demonstrate the effectiveness of Shy-hunyuan-MT. The model achieves consistent top-tier performance with an AutoRank score of 1.0 across all evaluated directions, while excelling on multiple automatic metrics, including XCOMET-XL, CometKiwi-XL, and GEMBA-ESA variants. Notably, the model demonstrates its ability to handle both high-resource and extremely low-resource language pairs with comparable proficiency, highlighting its strong generalization capabilities and effectiveness in cross-lingual transfer learning. The success of our approach validates several key design decisions: (1) the importance of domain-specific continual pre-training in adapting general-purpose LLMs for translation tasks, (2) the effectiveness of incorporating multiple COMET-based metrics as reward signals during policy optimization, and (3) the value of progressive training paradigms in building robust multilingual systems. These findings provide valuable insights for future research in neural machine translation and cross-lingual model adaptation. Looking forward, Shy-hunyuan-MT establishes a strong foundation for advancing multilingual translation technology, particularly in scenarios involving diverse language families and resource constraints. The framework's flexibility and consistent performance across varied linguistic contexts position it as a promising solution for real-world translation applications and a solid baseline for future improvements in the field.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Christian Buck and Philipp Koehn. 2016. Findings of the wmt 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli-Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The university of edinburgh's english-german and english-hausa submissions to the wmt21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Yuval Kirstain, Patrick Lewis, Sebastian Riedel, and Omer Levy. 2021. A few more examples may be worth billions of parameters. *arXiv preprint arXiv:2110.04374*.

Zheng Li, Mao Zheng, Mingyang Song, and Wenjie Yang. 2025. Tat-r1: Terminology-aware translation with reinforcement learning and word alignment. *Preprint*, arXiv:2505.21172.

Baohao Liao, Shahram Khadivi, and Sanjika Hewavitharana. 2021. Back-translation for large-scale multilingual machine translation. *CoRR*, abs/2109.08712.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Mingyang Song, Mao Zheng, Zheng Li, Wenjie Yang, Xuan Luo, Yue Pan, and Feng Zhang. 2025. Fastcurl: Curriculum reinforcement learning with stage-wise context scaling for efficient training r1-like reasoning models. *Preprint*, arXiv:2503.17287.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Kocmi Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2023. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *WMT23-Eighth Conference on Machine Translation*, pages 198–216.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Wenjie Yang, Mao Zheng, Mingyang Song, Zheng Li, and Sitong Wang. 2025. Ssr-zero: Simple self-rewarding reinforcement learning for machine translation. *Preprint*, arXiv:2505.16637.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

From SALAMANDRA to SALAMANDRATA: BSC Submission for WMT25 General Machine Translation Shared Task

Javier Garcia Gilabert[*1] **Xixian Liao**[*1] **Severino Da Dalt**[1] **Ella Bohman**[1] **Audrey Mash**[1] **Francesca De Luca Fornaciari**[1] **Irene Baucells**[1] **Joan Llop**[1] **Miguel Claramunt Argote**[1] **Carlos Escolano**[1,2] **Maite Melero**[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya

Abstract

In this paper, we present the SALAMANDRA**TA** family of models, an improved iteration of SALAMANDRA LLMs (Gonzalez-Agirre et al., 2025) specifically trained to achieve strong performance in translation-related tasks for 38 European languages. SALAMANDRA**TA** comes in two scales: 2B and 7B parameters. For both versions, we applied the same training recipe with a first step of continual pre-training on parallel data, and a second step of supervised fine-tuning on high-quality instructions.

The BSC submission to the WMT25 General Machine Translation shared task is based on the 7B variant of SALAMANDRA**TA**. We first adapted the model vocabulary to support the additional non-European languages included in the task. This was followed by a second phase of continual pre-training and supervised fine-tuning, carefully designed to optimize performance across all translation directions for this year's shared task. For decoding, we employed two quality-aware strategies: Minimum Bayes Risk Decoding and Tuned Re-ranking using COMET and COMET-KIWI respectively.

We publicly release both the 2B and 7B versions of SALAMANDRA**TA**, along with the newer SALAMANDRA**TA**-V2 model, on Hugging Face[1].

1 Introduction

Traditionally, Massively Multilingual Neural Machine Translation (MMNMT) relied on the encoder-decoder architecture to translate across multiple languages (Fan et al., 2021; NLLB Team et al., 2022). More recently, however, Large Language Models (LLMs) have demonstrated strong MMNMT capabilities (Zhu et al., 2024) and thus some works have proposed several strategies to improve the translation capabilities of a pre-trained LLM model and better align it with human translations (Zhang et al., 2023; Alves et al., 2024; Xu et al., 2024).

One such approach is continual pre-training using a combination of monolingual and parallel corpora followed by supervised fine-tuning (Alves et al., 2024). However, most previous approaches have predominantly relied on English-centric parallel corpora. This has been shown to bias the models towards English-centric latent representations (Zhang et al., 2025) which has been attributed to the language distribution used in the training corpora (Zhong et al., 2024). It is well known that training with only a single bridge language can negatively impact translation performance across zero-shot language pairs, due to limited cross-lingual transfer (Arivazhagan et al., 2019). Unlike previous works, in this paper we rely on parallel corpora only for the continual pre-training stage pivoting on three bridge languages.

When working with pre-trained language models on languages not covered by their original tokenizer, a highly effective solution involves replacing the existing tokenizer with a more comprehensive one that supports such languages. For the newly introduced tokens, embeddings must be initialized. In our work, these new embeddings were initialized to the average of all existing embeddings and then rapidly optimized through continual pre-training (CPT). This method has not only proven to be viable but also demonstrably improves the model's overall performance in the target languages, even if the original model was never exposed to data from these languages during its initial training (Da Dalt et al., 2024).

Throughout this paper, we present the SALAMANDRA**TA** family of models, which serve as the backbone models of the BSC team's submission to the WMT25 General Machine Translation Shared Task. Our participation covers 15 out of the 16

[*]Core Contributor.
[1]SALAMANDRA**TA**7B-V1 , SALAMANDRA**TA**2B-V1 and SALAMANDRA**TA**7B-V2 .

Proceedings of the Tenth Conference on Machine Translation, pages 614–637
November 8-9, 2025

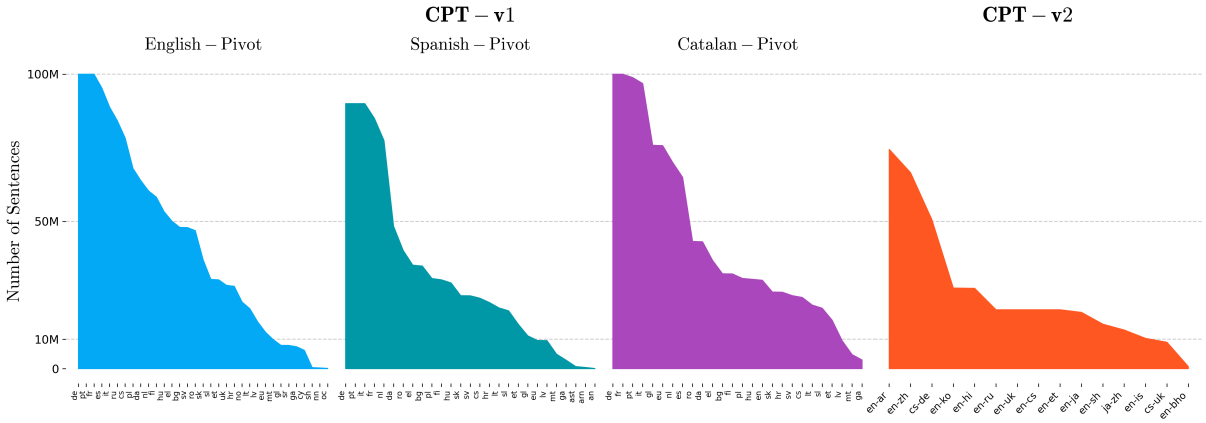


Figure 1: Distribution of sentence pairs for continual pre-training. The first three plots (■ **CPT-v1**) show the number of sentence pairs pivoting in English, Spanish and Catalan, respectively. The fourth plot (■ **CPT-v2**) corresponds to the second continual pre-training phase with direct language pairs.

translation directions in the general MT task under the constrained track. Additionally, we took part in the multilingual subtask for 7 out of the 16 directions. Contributions of this work are listed as follows:

- While most previous work have relied on English-centric parallel corpora for building translation-focused LLMs, we build SALAMANDRA**TA** pivoting in three languages for continual pre-training; English, Spanish and Catalan across 172 supervised directions.
- We show that instruction tuning improves both translation quality and robustness to character-level noise.
- We release all model checkpoints to facilitate reproducibility and future research on massively multilingual machine translation.

2 Data

Our base models are SALAMANDRA-2B and SALAMANDRA-7B (Gonzalez-Agirre et al., 2025), which were trained from scratch on highly multilingual data. However, SALAMANDRA models were not exposed to parallel data during pre-training. To address this, and following Alves et al. (2024), we improve their multilingual machine translation capabilities by performing continual pre-training on parallel data covering 38 European languages (35 of which were already present in the original pre-training corpus). This step is followed by supervised fine-tuning using high-quality instruction data. In this section, we detail the datasets used for both continual pre-training and supervised fine-tuning.

2.1 Continual pre-training

To train the SALAMANDRA**TA** models, we first compile a parallel corpus from publicly available data sources. A comprehensive list of these sources and the corresponding language pairs can be found in Table 5. We build two separate training sets: **CPT-v1** and **CPT-v2**. All data undergo initial filtering using LABSE (Feng et al., 2022), and off-target translations are excluded using the Lingua[2] library. After filtering, the data is de-duplicated and punctuation is normalized with the `Bifixer` library (Ramírez-Sánchez et al., 2020). The final corpora are formatted using the prompt template provided in Appendix Figure 3. Additional dataset details are available in Appendix C.

Using **CPT-v1** we continue pre-training SALAMANDRA 2B and 7B with the causal language modeling objective resulting in SALAMANDRA**TA**2B-BASE and SALAMANDRA**TA**7B-BASE models. Then, we use **CPT-v2** to continue pre-training SALAMANDRA**TA**7B-BASE.

■ **CPT-v1**: The first corpus, is employed during the initial round of continual pre-training (CPT), with the objective of enhancing the machine translation capabilities of SALAMANDRA across European languages. The final dataset has 38 languages across 6.57B sentence pairs and 172 machine translation directions in total pivoting in English, Spanish and Catalan, totaling in 424B tokens. We show in Figure 1 the data distribution of the **CPT-v1** corpus.

[2] https://github.com/pemistahl/lingua-py

■ **CPT-v2**: The second corpus, is used in the subsequent CPT round, where the focus shifts toward expanding coverage to include the additional language pairs featured in the WMT 2025 shared task. It includes 0.39B sentences across 14 languages and 15 directions, amounting to 27B tokens. To avoid the risk of catastrophic forgetting, we subsample 20M sentences for directions already present in **CPT-v1** (EN→CS, EN→ET, EN→RU, EN→UK). For EN→SH (English-to-Serbian, Latin script), we combined two sources from **CPT-v1**: English–Serbian (Latin script) data and English–Serbian (Cyrillic script) data, the latter converted to Latin script using rule-based transliteration. The per-direction data distribution is also shown in Figure 1. Note that we include the English-to-Hindi direction, which is not part of this year's shared task, in order to support better transfer for related languages such as Bhojpuri.

2.2 Instruction tuning

For instruction tuning we build two separate corpora: **IT-v1** and **IT-v2**. The first, **IT-v1**, is used to fine-tune SALAMANDRA**TA**2B-BASE and SALAMANDRA**TA**7B-BASE models into instruction-following models. The second corpus, is used to instruct SALAMANDRA**TA**7B-BASE after continue pre-training with **CPT-v2** corpus. We format each instruction using the `chatml` template (Open AI, 2023).

■ **IT-v1**: Following prior work on supervised fine-tuning for machine translation (Alves et al., 2024; Rei et al., 2024, 2025), we organize the instruction examples into three categories: pre-translation, translation, and post-translation tasks. The selection of tasks is motivated by the ablation results discussed in Section 4. The final corpus consists of 135k instructions, with the majority sourced from the TOWERBLOCKS collection (Alves et al., 2024). For translation related-tasks we focus on sentence, paragraph and document level data, primarily sourced from EUROPARL (Koehn, 2005). A big part of the data is drawn from multi-parallel datasets such as FLORES-200 (NLLB Team et al., 2022) or NTREX (Federmann et al., 2022), where a single source sentence has multiple translations in different target languages. When building the instruction tuning dataset, a naive strategy is to pivot trough different bridge languages across all languages including the complete dataset (e.g. for a given Catalan sentence that aligns to parallel sentences in, Spanish, French, and German, we might generate CA→ES, CA→FR, CA→DE and ES→CA, FR→CA, DE→CA). In our dataset we pivoted in five bridge languages: English, Catalan, Spanish, Basque and Galician across all the supported languages. However, this increases the number of duplicate training examples that share identical content on the target or source side. We found that doing this encourages target-side collapse, where the model produces off-target translations because many-to-one alignments blur the mapping between specific source inputs and their intended target languages. To mitigate this, we randomly sampled approximately equal numbers of translation instructions for each language pair. Further details on **IT-v1** are provided in Appendix C.

■ **IT-v2**: The second corpus, consisting of approximately 51k instructions, is constructed to focus on paragraph-level translation, context-aware machine translation, and sentence-level translation for the language directions included in the WMT 2025 shared task. To construct paragraph level data we source from FLORES-200-dev, NTREX and NEWSCOMMENTARY datasets. Similar to **IT-v1**, we applied random sampling when using multi-parallel datasets. In addition, we included data from TOWERBLOCKS that we considered relevant to our tasks. More details about **IT-v2** can be found in Appendix C.

3 SalamandraTA Models

The SALAMANDRA**TA** family is composed of two base models, 2B and 7B parameters, which were continually pre-trained on the **CPT-v1** corpus and subsequently instruction-tuned on **IT-v1**. For our submission to the WMT25 General Translation Shared Task, we further adapted the 7B model, resulting in SALAMANDRA**TA**-V2.

3.1 Adding WMT languages: SALAMANDRATA-V2

To expand the language coverage of SALAMANDRA**TA** and accommodate the additional languages required by the WMT25 General Translation Shared Task, we implemented vocabulary adaptation. We trained a new tokenizer on a corpus comprising the original languages augmented with monolingual text for the new languages not included in the original SALAMANDRA tokenizer: Chinese, Korean, Japanese, Arabic, and Bhojpuri.

The old tokenizer was replaced with the new one, which required re-initializing the embedding and unembedding layers. To address this, we modified these layers to ensure that tokens common to both the old and new tokenizers retained their original embeddings. The embeddings for the remaining, newly introduced tokens were initialized as the average of all existing embeddings. We expected this strategy to be particularly successful given that the two tokenizers share over 58% of their vocabulary. Figure 7 shows the fertility per language pair, comparing our new SALAMANDRA tokenizer against previous tokenizer, MADLAD400 and NLLB. On average, SALAMANDRA achieves a fertility of 1.88, outperforming both NLLB (2.00) and MADLAD400 (2.33) on WMT25 language pairs.

The subsequent section details the continual pre-training stage of our model. This stage aims not only to enhance the model's translation capabilities but also to recover the embeddings of these newly initialized tokens. More details can be found in Appendix D.

3.2 Model training

3.2.1 Continual pre-training

For this phase, we chose SALAMANDRA-2B and SALAMANDRA-7B as base models, using checkpoints preceding the annealing phase described in Gonzalez-Agirre et al. (2025). This choice was intentional: the annealing phase narrows the data sources to shape the model into a general-purpose downstream performer, which we considered misaligned with (or even counterproductive to) our goal of improving translation capabilities. The training strategy followed a schedule similar to that of the annealing phase. The learning rate was linearly warmed up over the first 2,000 steps, reaching a peak of 3e-5, and then decayed using a cosine schedule down to 3e-6. To mitigate the risk of exploding gradients, we applied gradient clipping with a maximum norm of 1.0 after the warm-up stage. We used NVIDIA NeMo as the training framework, and all other training hyperparameters were kept consistent with those used in the original SALAMANDRA pre-training (see Appendix E for more details). We trained the 7B model for 105k steps and the 2B model for 50k steps on the **CPT-v1** corpus tokenized with the original SALAMANDRA tokenizer (see Appendix Figure 10).

After vocabulary adaptation, we continually pre-train the resulting SALAMANDRA**TA**-7B model using **CPT-v2**. The training strategy followed the same training configuration as previously described.

3.2.2 Supervised Fine-tuning

We fine-tune SALAMANDRA**TA** base models using `FastChat` framework (Zheng et al., 2023). Hyperparameter details are provided in Appendix Table 10.

3.3 Evaluation

Metrics We assess translation quality using several metrics. For reference-based evaluation, we report scores from the learned metrics COMET (Rei et al., 2022a), BLEURT (Sellam et al., 2020), and METRICX (Juraska et al., 2023). For reference-free quality estimation (QE), we use COMET-KIWI (Rei et al., 2022b), and METRICX-QE. We also report two lexical-based metrics: CHRF (Popović, 2015) and BLEU (Papineni et al., 2002).

Datasets We used the FLORES-200-devtest dataset for ablation studies on the SALAMANDRA**TA** models. For evaluating translation quality on the WMT 2025 directions, we primarily relied on the WMT24++ dataset (Deutsch et al., 2025). An exception is the English to Bhojpuri direction, which is not included in WMT24++; for this case, we used FLORES-200-devtest for evaluation.

Baselines We compare the different SALAMANDRA**TA** variants against the translation LLM TOWER-V2 7B (Rei et al., 2024), as well as dedicated MMNMT models such as MADLAD400 7B (Kudugunta et al., 2023) and NLLB 3.3B (NLLB Team et al., 2022).

Decoding strategies For inference with the baseline, base, and instruction-tuned models, we employ beam search with a beam size of 5. Additionally, we experiment with two alternative decoding approaches: we use diverse beam search (Vijayakumar et al., 2018), which promotes output diversity by penalizing similar beams, and two post-decoding strategies applied to the generated candidates: Tuned Re-ranking Decoding (TRR) and Minimum Bayes Risk Decoding (MBR) (Eikema and Aziz, 2020) using the `mbrs` library (Deguchi et al., 2024). For diverse beam search we set a beam size of 20 and 5 beam groups. For post-decoding methods, we use COMET-22 as the quality metric for MBR and COMET-KIWI for TRR.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	71.7	-	79.7	-	-	-	-	81.9	-	**84.1**	76.8	-	-
MADLAD400 7B	82.7	83.2	76.8	-	82.1	71.1	72.4	73.7	81.7	78.3	81.8	82.8	76.4
NLLB 3.3B	79.5	80.4	76.6	-	78.3	70.1	72.7	70.3	77.9	80.3	76.9	78.9	68.4
SALAMANDRA**TA**2B													
BASE + CPT-V1	80.3	80.1	76.0	-	69.6	-	-	-	-	-	80.1	57.0	-
+ INSTRUCT-V1	80.7	80.3	76.5	-	78.0	-	-	-	-	-	76.0	78.0	-
+ TRR	84.3	86.0	80.5	-	83.3	-	-	-	-	-	80.4	81.8	-
+ MBR	85.6	87.0	81.4	-	84.0	-	-	-	-	-	81.5	83.5	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	81.9	79.8	76.6	-	78.0	-	-	-	-	-	81.5	82.2	-
+ INSTRUCT-V1	85.3	86.6	80.3	-	83.8	-	-	-	-	-	81.6	83.4	-
+ TRR	85.9	87.6	82.0	-	85.0	-	-	-	-	-	81.3	84.0	-
+ MBR	**87.2**	**88.7**	**82.9**	-	85.9	-	-	-	-	-	**82.6**	**85.1**	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	81.1	79.3	76.2	79.4	77.0	69.3	70.6	74.7	75.5	75.9	81.5	82.5	77.3
+ INSTRUCT-V2	83.1	85.3	79.3	83.9	84.1	77.4	71.3	81.1	80.9	80.2	80.4	82.3	77.8
+ TRR	85.3	87.3	81.8	84.9	85.1	79.7	74.2	82.7	83.3	82.5	81.3	84.2	79.6
+ MBR	86.6	88.5	82.4	86.3	**86.1**	**80.7**	**75.5**	**83.4**	**84.1**	83.6	82.5	**85.1**	**80.4**

Table 1: COMET scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models, and finally with the application of quality-aware decoding strategies (TRR and MBR). Using Minimum Bayes Risk (MBR) decoding consistently yields the best results.

4 Results

Table 1 presents the main translation quality results on the WMT24++ test set, measured in COMET scores for the language directions in the general MT task. We report extra metrics in Appendix F. We additionally evaluate SALAMANDRA**TA**-2B and SALAMANDRA**TA**-7B using COMET and METRICX for the language directions present in the multilingual subtask and report them in Appendix Table 17.

As shown in Table 1, instruction tuning yields significant gains over the CPT baselines, improving the SALAMANDRA**TA**-7B, SALAMANDRA**TA**-2B, and SALAMANDRA**TA**-V2 models by an average of 3.51, 4.40, and 3.60 COMET points, respectively.

Although further adapting the SALAMANDRA**TA**-7B model to WMT-2025 language pairs initially causes an average performance drop of 1.09 COMET points on the language directions shared between SALAMANDRA**TA**-7B and SALAMANDRA**TA**-V2, this gap is largely mitigated when employing quality-aware decoding strategies. Applying Minimum Bayes Risk (MBR) and Tuned Re-ranking (TRR) decoding strategies reduces this drop to 0.16 and 0.20 COMET points, respectively.

On the impact of adding non-MT-Tasks To better understand the impact of different instruction types on translation quality, we conduct an ablation study of instruction fine-tuning across four main task categories: machine translation (MT), pre-translation tasks (Pre-MT) (e.g., Named Entity Recognition), post-translation tasks (Post-MT) (e.g., Gender Bias Mitigation), and chat/code-related tasks[3]. Table 2 presents the model's performance after fine-tuning on each of these categories.

Instruction fine-tuning using MT tasks consistently yields the best overall performance across most evaluation metrics, with the exception of METRICX. For METRICX, a combination of MT, Pre-MT, and Post-MT instructions results in slightly improved performance. In contrast, adding only Pre-MT or Post-MT instructions shows no significant difference compared to the MT-only baseline. Incorporating Chat and Code instructions,

[3]This last group includes TOWERBLOCKS synthetic chat data and code instruction data.

	en→xx			xx→en		
	COMET	METRICX	BLEU	COMET	METRICX	BLEU
SALAMANDRA**TA**7B BASE + CPT-V1	0.85	1.73	34.60	**0.88**	1.15	44.22
Supervised Finetuning						
MT	**0.87**	1.33	**36.71**	**0.88**	1.17	**45.02**
+ Pre-MT + Post-MT	**0.87**	**1.14**	36.42	**0.88**	**1.09**	45.00
+ Chat + Code	**0.87**	1.36	35.58	**0.88**	1.16	44.81
MT + Post-MT	**0.87**	1.33	36.57	**0.88**	1.15	44.88
MT + Pre-MT	**0.87**	1.33	36.34	**0.88**	1.16	44.67

Table 2: Ablation study on the impact of different supervised fine-tuning tasks for the SALAMANDRA**TA**7B-BASE model. We report COMET, METRICX, and BLEU scores for English-to-Other (en→xx) and Other-to-English (xx→en) directions.

however, leads to a consistent drop in BLEU scores without measurable gains in other metrics.

Based on these findings, we concluded that for SALAMANDRA**TA**-2B and 7B, incorporating both Pre-MT and Post-MT tasks alongside MT tasks provided a slight benefit or at least no degradation in performance, leading to their inclusion in the **IT-v1** dataset. However, for SALAMANDRA**TA**-V2 which was specifically tailored for the WMT25 General Translation Shared Task, we made a deliberate choice to focus exclusively on MT instructions. While Pre-MT and Post-MT tasks might offer benefits, gathering high-quality, task-specific instruction data for the unique language pairs and domains present in WMT25 would have required significant additional effort beyond the scope of this work.

On the robustness to character noise Following Peters and Martins (2025), we investigate model robustness by injecting character-level noise into the source sentences of FLORES-200-devtest for the English to Spanish direction using adjacent swaps, duplications, and deletions at different noise levels. Figure 2 shows the relative degradation in BLEU score compared to zero-noise baseline. The SALAMANDRA**TA** 7B instruction-tuned model consistently shows greater resilience than the base model across all perturbation types. At the maximum noise level (1.0), the performance degradation of the instruction-tuned model is smaller by 17.63 p.p. for swaps, 20.61 p.p. for duplications, and 18.33 p.p. for deletions. These results demonstrate that instruction tuning effectively improves a model's robustness to character-level input corruptions.

Adding a low-resource language: The case of Bhojpuri Table 3 presents our ablation experiments for English to Bhojpuri translation direction. We find that during CPT, removing the EN→HI parallel data causes performance to drop from 9.32 to 0.35 BLEU and from 35.43 to 9.83 CHRF. This result provides clear evidence that the model relies on cross-lingual transfer from Hindi for translating to Bhojpuri. Finally, supervised fine-tuning (IT-V2) improves performance, improving the scores to 11.67 BLEU and 37.75 CHRF. This result shows the effectiveness of fine-tuning on high-quality data in the final stage, even for low-resource language pairs.

	BLEU	CHRF
Continual pre-training		
CPT-V2	9.32	35.43
CPT-V2 (no EN→HI)	0.35	9.83
Supervised Finetuning		
CPT-V2 + IT-V2	11.67	37.75

Table 3: Ablation results for English→Bhojpuri translation in terms of BLEU and CHRF on FLORES-200-devtest. The table compares the impact of removing the EN→HI direction from the CPT data and the effect of supervised fine-tuning (**IT-V2**).

5 Submission

For our WMT25 general and multilingual MT tasks submissions, we apply a chunking strategy, splitting each input instance at `\n\n` delimiter prior to translation. We made two submissions using

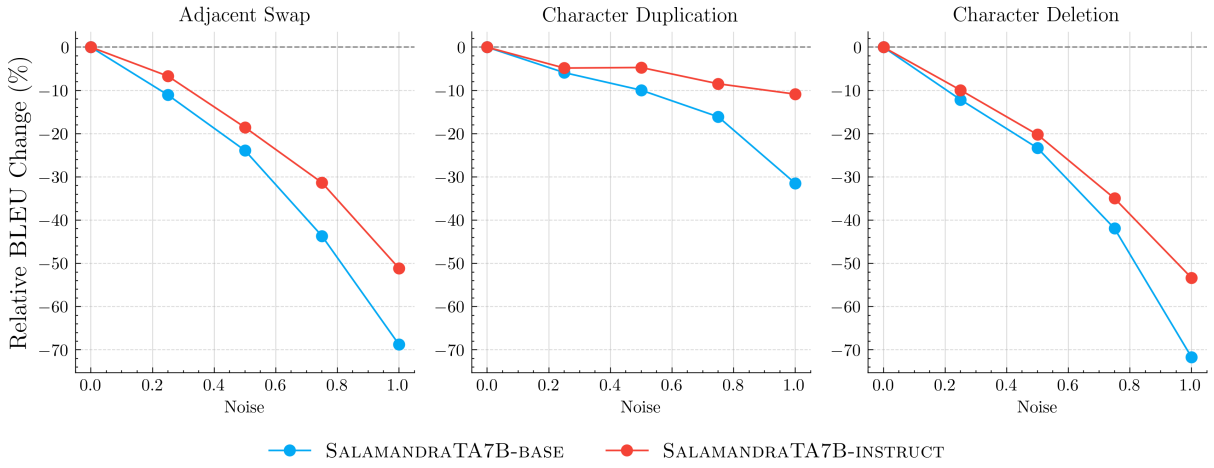


Figure 2: Relative change in BLEU scores (%) under increasing levels of input noise for three types of character-level perturbations: Adjacent Swap, Character Duplication, and Character Deletion.

two quality-aware decoding strategies: Minimum Bayes Risk Decoding employing COMET and Tuned Re-ranking relying on COMET-Kiwi.

6 Conclusion

In this paper, we introduced the SALAMANDRA**TA** family of models, a series of powerful, translation LLMs in 2B and 7B scales. Our approach combines a multi-stage training recipe, beginning with continual pre-training on parallel data that pivots through three languages: English, Spanish, and Catalan. This is followed by an instruction tuning stage to align the models with human translation outputs. For our WMT25 submission, we adapted our 7B model to new, non-European languages through vocabulary adaptation and a further round of continual pre-training and supervised fine-tuning.

Our experimental results show that instruction tuning is a critical step which not only improves translation quality but also the model's robustness against character-level noise. Furthermore, our analysis of the English-to-Bhojpuri direction validates the importance of including related languages during pre-training to enable cross-lingual transfer to low-resource pairs.

While our work successfully specializes models for translation and translation-related tasks, we observed that incorporating Chat and Code instructions during the supervised fine-tuning stage leads to a significant drop in translation quality as measured by BLEU. Future work could explore methods to mitigate this trade-off to train machine translation models that can follow general instructions without compromising their specialized translation capabilities.

7 Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work has been supported by the Spanish project PID2021-123988OB-C33 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is partially supported by MLLM4TRA (PID2024-158157OB-C32) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of ILENIA Project with reference 2022/TL22/00215337.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 303–304, Ghent, Belgium. European Association for Machine Translation.

CASMACAT. 2018. Global Voices Parallel Corpus 2018Q4. Accessed: July, 2025.

Severino Da Dalt, Joan Llop, Irene Baucells, Marc Pamies, Yishi Xu, Aitor Gonzalez-Agirre, and Marta Villegas. 2024. FLOR: On the effectiveness of language adaptation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7377–7388, Torino, Italia. ELRA and ICCL.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2024. mbrs: A library for minimum Bayes risk decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 351–362, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. XLEnt: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

ELRC-Share. 2020. Bilingual corpus made out of pdf documents from the european medicines agency (emea). .

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

878–891, Dublin, Ireland. Association for Computational Linguistics.

Aarón Galiano-Jiménez, Felipe Sánchez-Martínez, and Juan Antonio Pérez-Ortiz. 2024. PILAR: A collection of low-resource language corpora from the iberian peninsula. https://github.com/transducens/PILAR.

Pablo Gamallo, Marcos Garcia, Iria de-Dios-Flores, José Ramom Pichel Campos, Sandra Rodríguez Rey, and Daniel Bardanca. 2023a. NÓS corpus: Authentic English–Galician Parallel Corpus. Zenodo, https://doi.org/10.5281/zenodo.7675110. Accessed: July 2025.

Pablo Gamallo, Marcos García, Iria de Dios-Flores, José Ramom Pichel Campos, Sandra Rodríguez Rey, and Daniel Bardanca. 2023b. Nós corpus: Synthetic English–Galician Parallel Corpus. Zenodo, https://doi.org/10.5281/zenodo.7685180. Accessed: July 2025.

Mercedes García-Martínez, Laurent Bié, Aleix Cerdà, Amando Estela, Manuel Herranz, Rihards Krišlauks, Maite Melero, Tony O'Dowd, Sinead O'Gorman, Marcis Pinnis, Artūrs Stafanovič, Riccardo Superbo, and Artūrs Vasiļevskis. 2021. Neural translation for European Union (NTEU). In *Proceedings of Machine Translation Summit XVIII: Users and Providers Track*, pages 316–334, Virtual. Association for Machine Translation in the Americas.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *Preprint*, arXiv:2502.08489.

Kenneth Heafield, Elaine Farrow, Jelmer van der Linde, Gema Ramírez-Sánchez, and Dion Wiggins. 2022. The EuroPat corpus: A parallel corpus of European patent data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 732–740, Marseille, France. European Language Resources Association.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Open AI. 2023. [link].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ben Peters and Andre Martins. 2025. Did translation models get more robust without anyone Even noticing? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the*

Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Project Ilenia. 2024. GAITU Corpus: Catalan–Basque Synthetic Parallel Sentences. Hugging Face dataset, published approx. Dec 2024. Accessed: 2025-07-20; license: CC BY-NC-SA 4.0.

Projecte Aina-Language Technologies Unit, BSC. 2024. CA–EN Parallel Corpus: Catalan–English Synthetic Parallel Sentences. Hugging Face dataset, DOI:10.57967/hf/1913. Accessed: 2025-07-20; license: CC BY 4.0.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the EU bookshop corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1850–1855, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlüter. 2012. DGT-TM: A freely available translation memory in 22 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 454–459, Istanbul, Turkey. European Language Resources Association (ELRA).

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

The GNOME Project. n.d. GNOME. Accessed: July, 2025.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search: Decoding diverse solutions from neural sequence models. *Preprint*, arXiv:1610.02424.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. Exploring translation mechanism of large language models. *Preprint*, arXiv:2502.11806.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *Preprint*, arXiv:2408.10811.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A CPT Template

This section presents the template used to prepare parallel data for continued pre-training. We used only one single template. Placeholders:

- { source }: source sentence
- { target }: target sentence
- { source_lang }: source language name
- { target_lang }: target language name

Template used for CPT

```
{ source_lang }: { source }
{ target_lang }: { target }
```

Figure 3: Template used to format parallel data for CPT.

B Prompt templates used to construct translation instructions

All templates used to construct instructions were adapted from TowerBlocks (Alves et al., 2024). Figure 4 shows an example of a template used for translation instructions in our **IT-v1** and **IT-v2** datasets.

C Dataset

C.1 Continual pre-training v1

The pre-training corpus for **CPT-v1** consists of 424 billion tokens of Catalan-centric, Spanish-centric, and English-centric parallel data, including all of the official European languages plus Catalan, Basque, Galician, Asturian, Aragonese and Aranese. It amounts to 6,574,251,526 parallel sentence pairs.

This highly multilingual corpus is predominantly composed of data sourced from OPUS (Tiedemann, 2012), with additional data taken from the NTEU Project (García-Martínez et al., 2021), Aina Project,[4] and other sources (see Table 5, and Table 4 shows the mapping between the BCP-47 language code and the language name). Where little parallel Catalan ↔ xx data could be found, synthetic Catalan data was generated from the Spanish

[4] https://projecteaina.cat/

Template used for IT

```
Translate the following text from { source_lang } to { target_lang }:
{ source_lang }: { source }
{ target_lang }: { target }
```

Figure 4: Example of a prompt template used to construct translation instructions for **IT-v1** and **IT-v2**.

side of the collected Spanish ↔ xx corpora using Projecte Aina's Spanish-Catalan model.[5] The final distribution of languages is shown in Figure 1.

Datasets with "-BSC" in their names (e.g., BOUA-SYNTH-BSC, DOGV-SYNTH-BSC) are synthetic datasets obtained by machine translating pre-existing monolingual corpora with our own seq-to-seq models. These datasets were generated internally for model training and are not published.

C.2 Continual pre-training v2

In **CPT-v2** we focused on the language pairs featured in the WMT 2025 shared task. For pairs involving European languages, we reused part of the data from **CPT-v1**. Specifically, we sampled 20M sentence pairs each for English–Czech, English–Estonian, and English–Russian from the **CPT-v1** data. For English–Serbian (Latin), we included the authentic English–Serbian (Latin) parallel dataset from **CPT-v1**. Additionally, we transliterated the Serbian side of the English–Serbian (Cyrillic) dataset into Latin script, taking advantage of the one-to-one correspondence between the two scripts. For English–Icelandic, Czech–Ukrainian, and Czech–German, we used the WMT 2025 Translation Task Training Data.[6]

For language pairs involving non-European languages, we used sentence-level data from the WMT 2025 Translation Task Training Data. The Chinese side of all datasets were first processed using the Hanzi Identifier to detect Traditional Chinese,[7] which was subsequently converted to Simplified Chinese using OpenCC.[8] We also included paragraph-level English–Arabic data by concatenating sentences from NEWSCOMMENTARY.

We created two versions of **CPT-v2**. The first included only the language pairs featured in the WMT25 shared task. In the second, we additionally included English–Hindi data from the OPUS

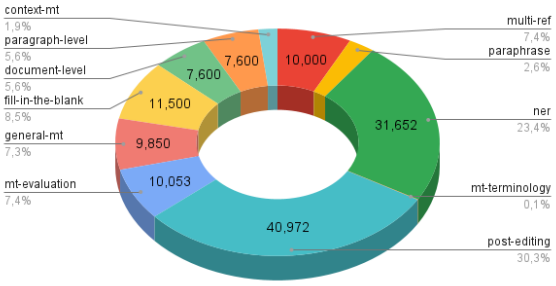


Figure 5: Distribution of tasks in **IT-v1** .

corpora CCMatrix (Schwenk et al., 2021b), MultiHPLT (de Gibert et al., 2024), NLLB (NLLB Team et al., 2022), and Samanantar (Ramesh et al., 2022), to support the model's performance on Bhojpuri (which uses the Devanagari script).

The pre-training corpus for **CPT-v2** without English-Hindi consists of 24 billion tokens, amounting to 366,179,935 parallel sentence pairs. For **CPT-v2** with English-Hindi, the corpus contains 26 billion tokens and 393,507,678 parallel sentence pairs. The data distribution is shown in Figure 1, and the corresponding sources are listed in Table 6.

As shown in Section 4, continual pre-training with Hindi data led to better performance, particularly for Bhojpuri.

C.3 Instruction tuning v1

During **IT-v1** the model was fine-tuned on ~135k instructions, primarily targeting machine translation performance for Catalan, English, and Spanish. Additional instruction data for other European and closely related Iberian languages was also included.

A portion of our fine-tuning data comes directly from, or is sampled from TOWERBLOCKS. While tasks related to machine translation are included, it is important to note that no chat data was used in the fine-tuning process. The final distribution of tasks is shown in Figure 5. The full list of tasks included in **IT-v1** is shown in Table 7.

[5] https://huggingface.co/projecte-aina/aina-translator-es-ca

[6] https://www2.statmt.org/wmt25/mtdata/

[7] https://github.com/tsroten/hanzidentifier

[8] https://github.com/BYVoid/OpenCC

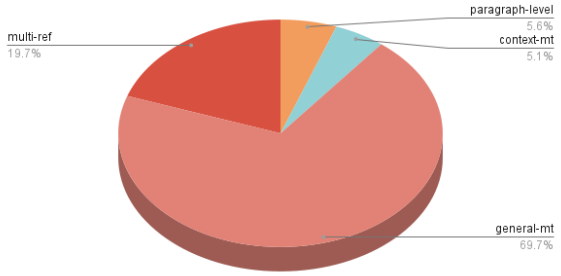


Figure 6: Distribution of tasks in **IT-v2** .

C.4 Instruction tuning v2

In **IT-v2** we focused on the languages pairs featured in the WMT 2025 shared task. We included paragraph-level data during instruction tuning to support paragraph-level translation. We constructed this data by concatenating adjacent sentences (randomly grouping 2, 3, or 4) from the same article or document in FLORES-200-dev, NTREX, and NEWSCOMMENTARY. To prevent over-representation of these sources, we sampled approximately equal amounts of paragraph-level data for each language pair. Serbian Cyrillic data from FLORES-200-dev was transliterated into Serbian Latin. In addition, we included data from TOWERBLOCKS that we considered relevant to our tasks. The instruction tuning dataset is summarized in Table 8 and the distribution of tasks is shown in Figure 6.

D Tokenizer

We evaluated the trained tokenizer using fertility metric on the FLORES-200 dataset (see Figure 7). For a given tokenizer T and a set of sentences S, fertility is defined as the ratio of the total number of tokens produced by T to the total number of words in S. Formally:

$$\text{Fertility}(T, S) = \frac{\#\text{tokens in } T(S)}{\#\text{words in } S} \quad (1)$$

The results in Figure 7 indicate that SALAMANDRA**TA**7B-V2 consistently achieves the lowest fertility scores on average among WMT25 languages.

E Training

F Results

Language Code	Language
ar	Arabic
arn	Aranese
ast	Asturian
arg	Aragonese
bho	Bhojpuri
bg	Bulgarian
ca	Catalan
cs	Czech
cy	Welsh
da	Danish
de	German
el	Greek
es	Spanish
en	English
et	Estonian
eu	Basque
fi	Finnish
fr	French
ga	Irish
gl	Galician
hi	Hindi
hr	Croatian
hu	Hungarian
is	Icelandic
it	Italian
ja	Japanese
ko	Korean
lt	Lithuanian
lv	Latvian
mt	Maltese
nl	Dutch
nn	Norwegian Nynorsk
no	Norwegian
oc	Occitan
pl	Polish
pt	Portuguese
ro	Romanian
ru	Russian
sh	Serbian (Latin)
sk	Slovak
sl	Slovenian
sr	Serbian (Cyrillic)
sv	Swedish
uk	Ukrainian
val	Catalan-Valencian
zh	Chinese

Table 4: Mapping from BCP-47 language codes to full language names.

Dataset	Ca-xx Languages	Es-xx Languages	En-xx Languages
AINA (Projecte Aina-Language Technologies Unit, BSC, 2024)	en		
ARANESE-SYNTH-CORPUS-BSC	arn		
BOUA-SYNTH-BSC		val	
BOUMH (Galiano-Jiménez et al., 2024)		val	
BOUA-PILAR (Galiano-Jiménez et al., 2024)		val	
CCMatrix (Schwenk et al., 2021b)	eu		ga
DGT (Steinberger et al., 2012)	bg, cs, da, de, el, et, fi, fr, ga, hr, hu, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	da, et, ga, hr, hu, lt, lv, mt, sh, sl	
DOGV-SYNTH-BSC		val	
DOGV-PILAR (Galiano-Jiménez et al., 2024)		val	
ELRC-EMEA (ELRC-Share, 2020)	bg, cs, da, hu, lt, lv, mt, pl, ro, sk, sl	et, hr, lv, ro, sk, sl	
EMEA (Tiedemann, 2012)	bg, cs, da, el, fi, hu, lt, mt, nl, pl, ro, sk, sl, sv	et, mt	
EUBookshop (Skadiņš et al., 2014)	lt, pl, pt	cs, da, de, el, fi, fr, ga, it, lv, mt, nl, pl, pt, ro, sk, sl, sv	cy, ga
Europarl (Koehn, 2005)		bg, cs, da, el, en, fi, fr, hu, lt, lv, nl, pl, pt, ro, sk, sl, sv	
Europat (Heafield et al., 2022)		en, hr	no
GAITU Corpus (Project Ilenia, 2024)			eu
KDE4 (Tiedemann, 2012)	bg, cs, da, de, el, et, eu, fi, fr, ga, gl, hr, it, lt, lv, nl, pl, pt, ro, sk, sl, sv	bg, ga, hr	cy, ga, nn, oc
GlobalVoices (CASMACAT, 2018; Tiedemann, 2012)	bg, de, fr, it, nl, pl, pt	bg, de, fr, pt	
GNOME (The GNOME Project, n.d.; Tiedemann, 2012)	eu, fr, ga, gl, pt	ga	cy, ga, nn
JRC-Arquis (Steinberger et al., 2006)	cs, da, et, fr, lt, lv, mt, nl, pl, ro, sv		et
LES-CORTS-VALENCIANES-SYNTH-BSC		val	
MaCoCu (Bañón et al., 2022)	en		hr, mt, uk
MultiCCAligned (El-Kishky et al., 2020)	bg, cs, de, el, et, fi, fr, hr, hu, it, lt, lv, nl, pl, ro, sk, sv	bg, fi, fr, hr, it, lv, nl, pt	bg, cy, da, et, fi, hr, hu, lt, lv, no, sl, sr, uk
MultiHPLT (de Gibert et al., 2024)	en, et, fi, ga, hr, mt	fi, ga, gl, hr, mt, nn, sr	
MultiParaCrawl (Bañón et al., 2020)	bg, da	de, en, fr, ga, hr, hu, it, mt, pt	bg, cs, da, de, el, et, fi, fr, ga, hr, hu, lt, lv, mt, nn, pl, ro, sk, sl, uk
MultiUN (Eisele and Chen, 2010)		fr	
News-Commentary (Tiedemann, 2012)		fr	
NLLB (NLLB Team et al., 2022)	bg, da, el, en, et, fi, fr, gl, hu, it, lt, lv, pt, ro, sk, sl	bg, cs, da, de, el, et, fi, fr, hu, it, lt, lv, nl, pl, pt, ro, sk, sl, sv	bg, cs, cy, da, de, el, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, no, oc, pl, pt, ro, ru, sk, sl, sr, sv, uk
NÓS Authentic Corpus (Gamallo et al., 2023a)			gl
NÓS Synthetic Corpus (Gamallo et al., 2023b)			gl
NTEU (García-Martínez et al., 2021)	bg, cs, da, de, el, en, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	da, et, ga, hr, lt, lv, mt, ro, sk, sl, sv	
OpenSubtitles (Lison and Tiedemann, 2016)	bg, cs, da, de, el, et, eu, fi, gl, hr, hu, lt, lv, nl, pl, pt, ro, sk, sl, sv	da, de, fi, fr, hr, hu, it, lv, nl	bg, cs, de, el, et, hr, fi, fr, hr, hu, no, sl, sr
OPUS-100 (Zhang et al., 2020; Tiedemann, 2012)	en		gl
StanfordNLP-NMT (Luong and Manning, 2016; Luong et al., 2015; Luong and Manning, 2015)			cs
Tatoeba (Tiedemann, 2012)	de, pt	pt	
TildeModel (Rozis and Skadiņš, 2017)	bg	et, hr, lt, lv, mt	
UNPC (Ziemski et al., 2016)		en, fr	ru
PILAR-VALENCIAN-AUTH (Galiano-Jiménez et al., 2024)		val	
PILAR-VALENCIAN-SYNTH (Galiano-Jiménez et al., 2024)		val	
WikiMatrix (Schwenk et al., 2021a)	bg, cs, da, de, el, et, eu, fi, fr, gl, hr, hu, it, lt, nl, pl, pt, ro, sk, sl, sv	bg, en, fr, hr, it, pt	oc, sh
Wikimedia			cy, nn
XLENT (El-Kishky et al., 2021)	eu, ga, gl	ga	cy, et, ga, gl, hr, oc, sh

Table 5: Data sources of **CPT-v1**.

Source	Language Pair
WMT 2025 Translation Task Training Data	en-ar en-zh cs-de en-ko en-ja ja-zh en-is cs-uk en-bho
NEWSCOMMENTARY (paragraph-level)	en-ar
CCMATRIX (Schwenk et al., 2021b)	en-hi
MULTIHPLT (de Gibert et al., 2024)	en-hi
NLLB (NLLB Team et al., 2022)	en-hi
SAMANANTAR (Ramesh et al., 2022)	en-hi
CPT-v1	en-cs en-et en-ru en-uk en-sh

Table 6: Data sources of **CPT-v2**.

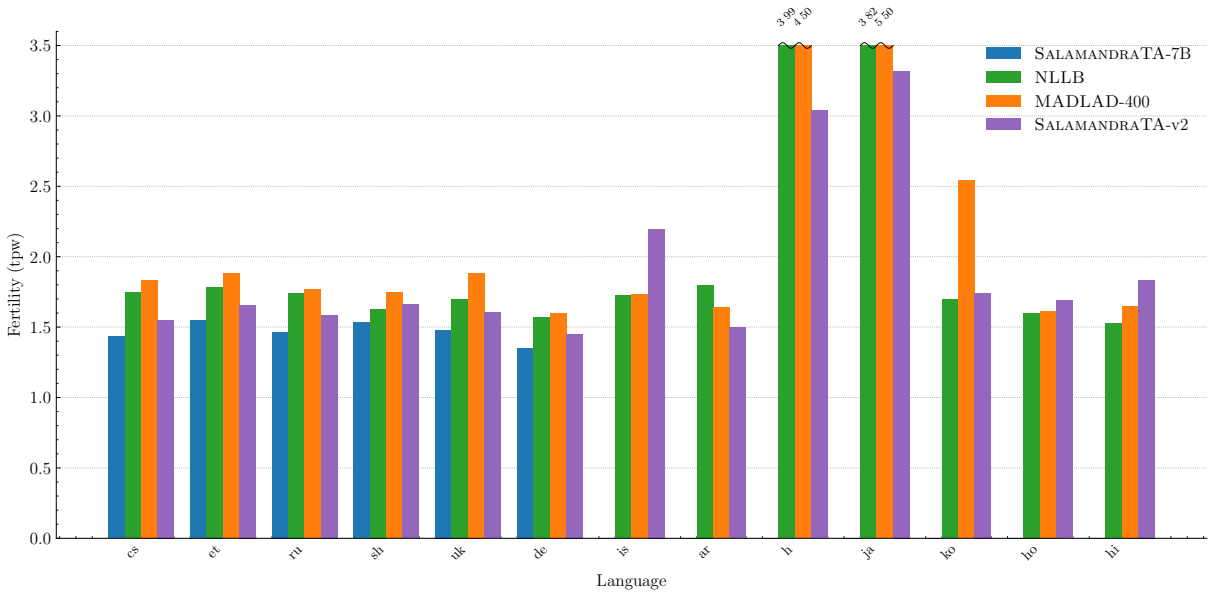


Figure 7: Tokenization fertility comparison across 13 languages from the FLORES-200 dataset. Fertility is shown on the vertical axis for each language on the horizontal axis. Results are presented for four multilingual models: SALAMANDRATA-7B, NLLB, MADLAD-400, and SALAMANDRATA-V2.

Category	Task	Source	Languages	Count
Pre-Translation	Named-entity Recognition	ANCORA-CA-NER	ca	12,059
		BASQUEGLUE, EUSIE	eu	4,304
		SLI NERC Galician Gold Corpus	gl	6,483
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	pt	854
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	nl	800
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	es	1,654
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	en	1,671
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	ru	800
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	it	858
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	fr	857
		TOWERBLOCKS: MULTICONER 2022-2023 Dev	de	1,312
Translation	Multi-reference Translation	TOWERBLOCKS: TATOEBA Dev	mixed	10,000
	Terminology-aware Translation	TOWERBLOCKS: WMT21 TERMINOLOGY DEV	en-ru	50
		TOWERBLOCKS: WMT21 TERMINOLOGY DEV	en-fr	29
	Fill-in-the-Blank	Non-public	Five pivot languages (ca, es, eu, gl, en) paired with European languages (cs, da, de, el, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv)	11,500
	General Machine Translation	TOWERBLOCKS: WMT14 to WMT21, NTREX, FLORES DEV, FRMT, QT21, APEQUEST, OPUS (Quality Filtered), MT-GENEVAL	nl-en, en-ru, it-en, fr-en, es-en, en-fr, ru-en, fr-de, en-nl, de-fr	500
		FLORES DEV, NTREX	Four pivot languages (es, ca, eu, gl) paired with the rest of languages. We sample 50 instances for each pair.	9350
	Document-level Translation	Non-public	Two pivot languages (es, en) paired with European languages (bg, cs, da, de, el, et, fi, fr, hu, it, lt, lv, nl, pl, pt, ro, ru, sk, sv)	7,600
	Paragraph-level Translation	Non-public	Two pivot languages (es, en) paired with European languages (bg, cs, da, de, el, et, fi, fr, hu, it, lt, lv, nl, pl, pt, ro, ru, sk, sv)	7,600
	Context-Aware Translation	TOWERBLOCKS: MT-GENEVAL	en-it	348
			en-ru	454
			en-fr	369
			en-nl	417
			en-es	431
			en-de	558
Post-Translation	Paraphrase	TOWERBLOCKS: PAWS-X DEV	mixed	3,521
	Machine Translation Evaluation	TOWERBLOCKS (sample): WMT20 to WMT22 METRICS MQM, WMT17 to WMT22 METRICS DIRECT ASSESSMENTS	en-ru, en-pl, ru-en, en-de, en-ru, de-fr, de-en, en-de	353
		Non-public	Four pivot languages (eu, es, ca, gl) paired with European languages (bg, cs, da, de, el, en, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv)	9,700
	Automatic Post Editing	TOWERBLOCKS: QT21, APEQUEST	en-fr	6,133
		TOWERBLOCKS: QT21, APEQUEST	en-nl	9,077
		TOWERBLOCKS: QT21, APEQUEST	en-pt	5,762
		TOWERBLOCKS: QT21, APEQUEST	de-en	10,000
		TOWERBLOCKS: QT21, APEQUEST	en-de	10,000
Total				**135,404**

Table 7: Overview of tasks, data sources, language coverage, and counts in **IT-v1**.

Category	Task	Source	Languages	Count
Translation	Paragraph-level Translation	FLORES DEV	en-ar	30
			en-bho	30
			en-ja	30
			en-uk	30
			en-ru	21
			cs-uk	30
			ja-zh	30
			en-zh	30
			en-ko	30
			en-et	30
			en-is	30
			en-sh	30
			en-cs	30
			cs-de	30
		NTREX	en-ja	58
			en-uk	58
			en-ru	50
			cs-uk	58
			ja-zh	58
			en-zh	58
			en-ko	58
			en-et	58
			en-is	58
			en-sh	58
			en-cs	58
			cs-de	58
		NEWS COMMENTARY	en-zh	250
			cs-de	250
			en-cs	250
			en-de	250
			en-ja	250
			ja-zh	250
			en-ru	250
	Context-Aware Translation	TOWERBLOCKS: MT-GENEVAL	en-it	348
			en-fr	369
			en-nl	417
			en-es	431
			en-de	558
			en-ru	454
	Multi-reference Translation	TOWERBLOCKS: TATOEBA Dev	mixed	10,000
	General Machine Translation	TOWERBLOCKS: WMT14 to WMT21, NTREX, FLORES DEV, FRMT, QT21, APE-QUEST, OPUS (Quality Filtered), MT-GENEVAL	en-ru	22,112
			en-zh	10,521
			en-ko	2,782
Total				**50,841**

Table 8: Overview of tasks, data sources, language coverage, and counts in **IT-v2**.

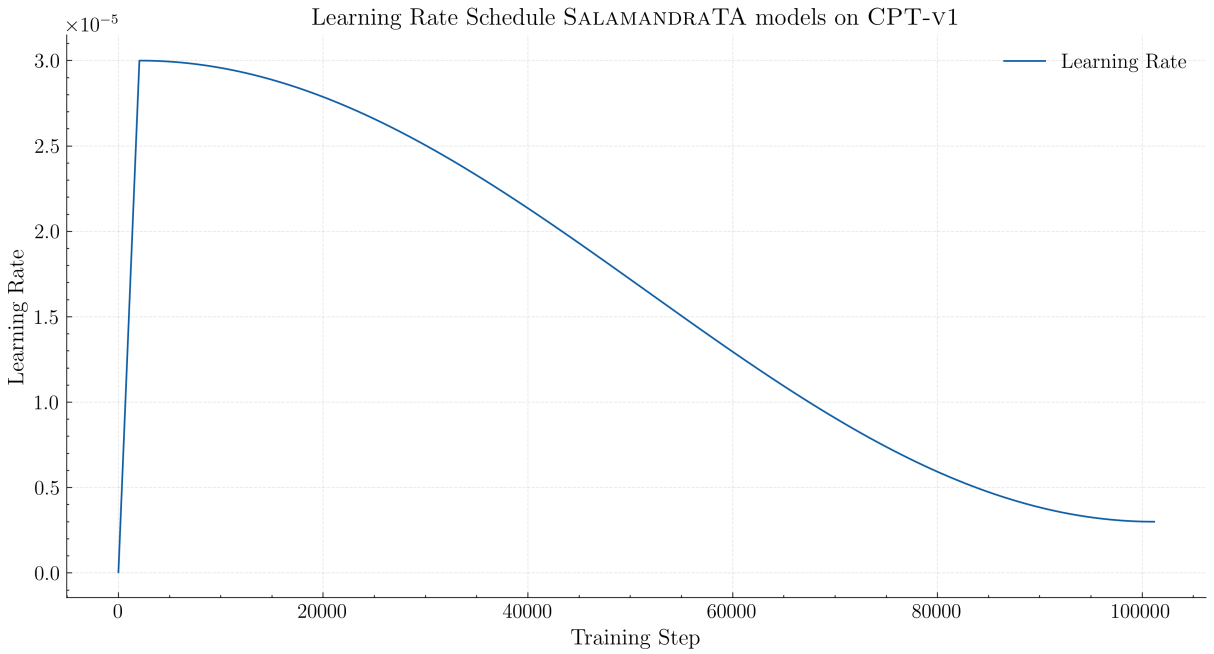


Figure 8: Learning Rate for the SALAMANDRA**TA**-7B and SALAMANDRA**TA**-2B on **CPT-v1**.

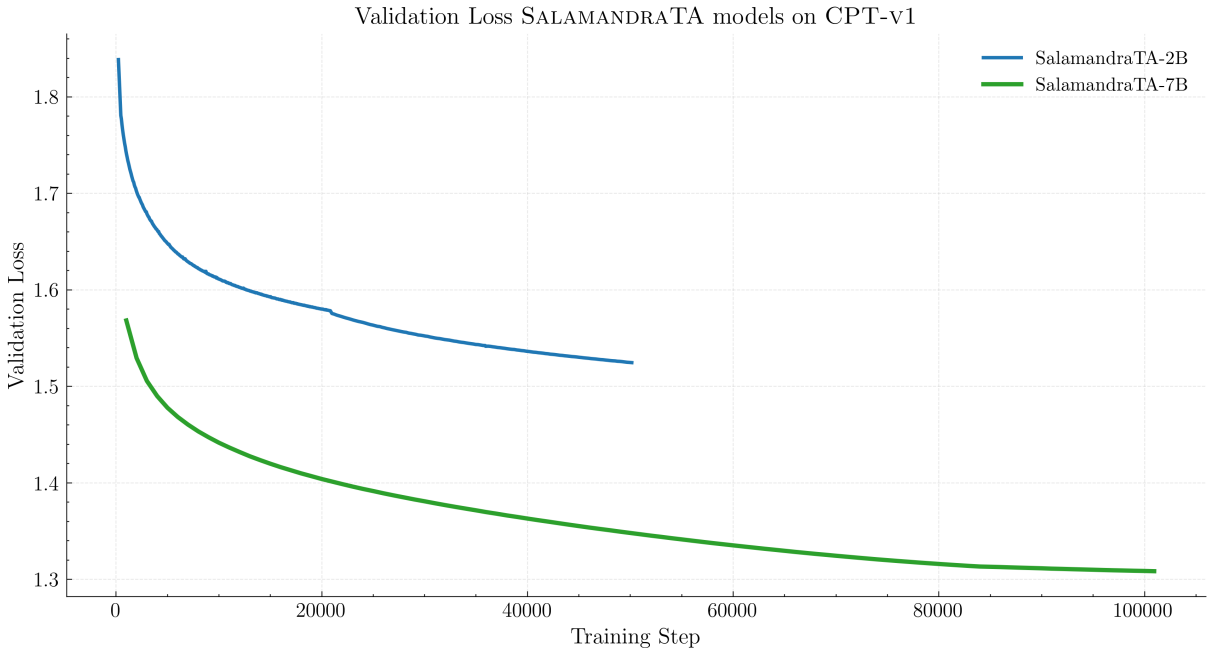


Figure 9: Validation loss for the SALAMANDRA**TA**-7B and SALAMANDRA**TA**-2B on **CPT-v1**.

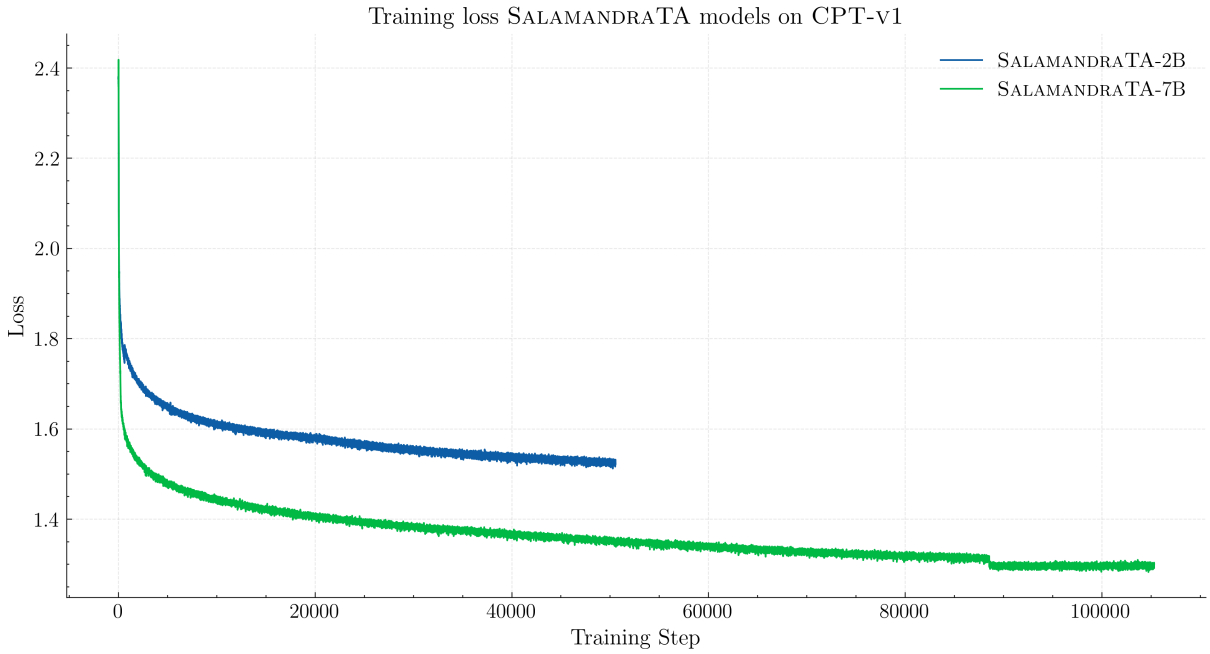


Figure 10: Training loss for the SALAMANDRA**TA**-7B and SALAMANDRA**TA**-2B on **CPT-v1**.

Table 9: Hyperparameters for SALAMANDRA**TA** continual pre-training.

Hyperparameter	**Value**
Micro Batch Size	2
Global Batch Size	512
Optimizer	Distributed Fused Adam
Learning Rate	3e-5
Minimum LR	3e-6
Weight Decay	0.1
Betas	(0.9, 0.95)
LR Scheduler	CosineAnnealing
Warmup Steps	2048
Mixed Precision	AMP O2
Sequence Length	8,192
Gradient Sync DType	bfloat16

Table 10: Hyperparameters for SALAMANDRA**TA** supervised-fine tuning.

Hyperparameter	**Value**
Train epochs	1
Train batch size per device	1
Gradient accumulation steps	16
Learning rate	1e-5
Weight decay	0
Warmup ratio	0.03
LR scheduler	Cosine
Model max length	8,192

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	11.1	-	21.2	-	-	-	-	35.6	-	24.7	18.4	-	-
MADLAD400 7B	28.4	27.2	22.5	-	26.8	17.5	6.9	30.2	19.8	25.3	25.2	20.9	20.8
NLLB 3.3B	23.0	21.8	20.7	-	23.4	16.2	6.9	23.9	13.6	22.5	19.4	16.4	15.5
SALAMANDRA**TA**2B													
BASE + CPT-V1	17.9	19.4	18.1	-	9.5	-	-	-	-	-	19.8	3.5	-
+ INSTRUCT-V1	17.1	12.1	13.7	-	14.6	-	-	-	-	-	10.2	10.7	-
+ TRR	24.7	23.5	19.4	-	24.9	-	-	-	-	-	20.5	17.5	-
+ MBR	25.1	22.1	19.4	-	24.6	-	-	-	-	-	21.1	17.4	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	25.9	25.1	20.2	-	25.6	-	-	-	-	-	24.9	20.1	-
+ INSTRUCT-V1	29.0	27.7	22.2	-	28.7	-	-	-	-	-	24.4	20.9	-
+ TRR	26.4	25.4	21.2	-	27.1	-	-	-	-	-	22.4	19.6	-
+ MBR	26.8	25.9	20.9	-	27.1	-	-	-	-	-	23.5	20.1	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	25.6	24.7	19.6	26.1	24.1	16.9	5.3	33.0	11.9	17.4	24.8	20.6	20.1
+ INSTRUCT-V2	27.3	25.7	19.5	27.8	29.2	17.6	6.0	36.6	14.4	18.8	20.0	19.1	22.3
+ TRR	26.5	25.1	21.0	26.6	26.7	17.4	6.1	35.8	17.7	20.9	22.6	20.3	22.3
+ MBR	26.1	25.4	20.4	27.0	27.5	17.5	6.3	36.2	16.7	20.9	22.7	20.6	22.1

Table 11: BLEU scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	39.6	-	49.7	-	-	-	-	32.5	-	32.1	49.2	-	-
MADLAD400 7B	55.0	57.8	49.7	-	53.2	43.4	36.2	27.7	28.0	31.5	54.7	47.8	20.6
NLLB 3.3B	49.7	51.7	46.6	-	48.6	40.9	35.9	22.4	23.6	29.6	47.7	42.8	15.9
SALAMANDRA**TA**2B													
BASE + CPT-V1	48.4	51.7	44.5	-	33.6	-	-	-	-	-	49.7	15.5	-
+ INSTRUCT-V1	49.3	47.6	44.9	-	45.9	-	-	-	-	-	44.7	40.4	-
+ TRR	52.7	55.7	48.6	-	52.3	-	-	-	-	-	51.4	45.5	-
+ MBR	52.5	55.0	48.4	-	51.9	-	-	-	-	-	51.8	46.0	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	52.8	55.6	48.7	-	52.3	-	-	-	-	-	54.0	47.9	-
+ INSTRUCT-V1	55.9	58.4	50.7	-	55.1	-	-	-	-	-	54.4	48.6	-
+ TRR	54.0	57.3	50.1	-	54.2	-	-	-	-	-	52.9	47.8	-
+ MBR	54.4	57.2	50.1	-	54.2	-	-	-	-	-	53.9	48.2	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	52.6	54.7	48.2	54.5	51.7	42.2	34.6	28.7	22.8	26.3	54.4	47.9	22.0
+ INSTRUCT-V2	53.9	56.8	48.7	56.8	54.9	43.8	35.5	32.7	26.9	28.5	52.2	47.2	21.1
+ TRR	54.3	57.2	50.0	56.0	54.2	44.6	36.2	32.5	28.2	28.8	53.2	48.4	21.7
+ MBR	54.0	57.3	49.6	56.5	54.4	44.4	36.3	32.8	27.9	28.9	53.7	48.4	21.6

Table 12: CHRF scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	6.69	-	4.16	-	-	-	-	3.83	-	3.70	2.25	-	-
MADLAD400 7B	4.28	4.14	5.50	-	4.18	7.18	7.75	6.60	4.49	5.98	1.73	4.04	6.05
NLLB 3.3B	5.95	6.03	6.38	-	6.64	8.46	7.71	7.91	6.09	5.74	2.74	6.45	8.12
SALAMANDRA**TA**2B													
BASE + CPT-V1	5.03	5.08	5.58	-	6.53	-	-	-	-	-	2.02	6.24	-
+ INSTRUCT-V1	3.99	4.19	4.90	-	5.28	-	-	-	-	-	2.40	5.10	-
+ TRR	3.02	2.67	3.86	-	3.82	-	-	-	-	-	1.63	3.91	-
+ MBR	3.02	2.82	3.83	-	3.92	-	-	-	-	-	1.63	3.85	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	4.43	5.24	5.30	-	5.58	-	-	-	-	-	1.73	4.12	-
+ INSTRUCT-V1	2.87	2.30	3.76	-	3.60	-	-	-	-	-	1.52	3.46	-
+ TRR	2.51	2.00	3.21	-	3.11	-	-	-	-	-	1.48	3.23	-
+ MBR	2.48	1.96	3.19	-	3.12	-	-	-	-	-	1.43	3.22	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	4.91	5.26	5.52	6.54	5.97	8.24	6.87	5.79	5.89	6.40	1.73	4.03	5.08
+ INSTRUCT-V2	3.60	2.79	4.13	4.44	3.44	5.26	8.48	4.03	4.59	5.15	1.77	3.83	4.66
+ TRR	2.81	2.08	3.30	3.96	2.98	4.43	7.47	3.60	3.98	4.50	1.50	3.25	4.13
+ MBR	2.79	2.12	3.35	4.00	2.99	4.57	7.73	3.62	4.00	4.51	1.49	3.28	4.21

Table 13: METRICX scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	55.9	-	61.7	-	-	-	-	61.8	-	59.8	62.4	-	-
MADLAD400 7B	69.3	71.2	58.7	-	62.4	52.4	39.2	53.2	53.8	52.8	68.9	61.5	54.5
NLLB 3.3B	65.2	67.7	58.3	-	60.4	51.0	40.3	48.5	45.9	53.2	62.9	58.6	43.8
SALAMANDRA**TA**2B													
BASE + CPT-V1	66.2	67.3	57.9	-	49.8	-	-	-	-	-	67.1	29.1	-
+ INSTRUCT-V1	68.5	69.5	59.7	-	61.7	-	-	-	-	-	66.1	59.4	-
+ TRR	70.7	73.3	62.4	-	64.5	-	-	-	-	-	68.0	61.2	-
+ MBR	70.7	73.1	61.9	-	64.4	-	-	-	-	-	68.3	62.6	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	68.3	67.1	58.2	-	60.5	-	-	-	-	-	68.5	63.2	-
+ INSTRUCT-V1	72.5	75.4	62.6	-	66.7	-	-	-	-	-	69.5	64.4	-
+ TRR	72.6	75.7	64.2	-	67.4	-	-	-	-	-	69.2	65.0	-
+ MBR	73.1	75.9	63.9	-	67.6	-	-	-	-	-	70.0	64.9	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	67.0	66.8	58.0	68.5	59.3	52.0	41.5	54.2	48.9	50.3	68.5	63.3	55.5
+ INSTRUCT-V2	69.8	74.1	62.2	73.3	67.6	58.4	38.5	61.4	54.1	55.5	69.0	64.6	55.7
+ TRR	71.8	75.2	63.8	73.7	67.7	59.2	39.5	62.6	55.6	56.8	69.6	65.3	57.1
+ MBR	71.8	75.4	63.8	74.1	67.8	59.2	39.2	62.4	55.6	56.8	69.5	65.4	56.8

Table 14: BLEURT scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	4.87	-	2.28	-	-	-	-	2.46	-	1.74	3.50	-	-
MADLAD400 7B	3.38	3.38	3.89	-	2.94	4.95	5.31	6.00	3.50	3.66	3.28	3.31	8.32
NLLB 3.3B	4.83	4.92	4.80	-	4.91	6.11	4.61	7.83	4.48	3.21	6.35	5.16	9.65
SALAMANDRA**TA**2B													
BASE + CPT-V1	3.73	4.02	3.46	-	5.48	-	-	-	-	-	3.97	4.46	-
+ INSTRUCT-V1	2.89	3.46	3.21	-	3.67	-	-	-	-	-	4.12	3.38	-
+ TRR	1.78	1.74	1.96	-	2.02	-	-	-	-	-	2.59	1.94	-
+ MBR	1.86	1.92	2.01	-	2.21	-	-	-	-	-	2.68	2.03	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	3.40	4.15	3.27	-	3.71	-	-	-	-	-	3.17	2.73	-
+ INSTRUCT-V1	1.82	1.75	2.07	-	2.14	-	-	-	-	-	2.69	1.87	-
+ TRR	1.49	1.42	1.63	-	1.69	-	-	-	-	-	2.46	1.58	-
+ MBR	1.52	1.44	1.74	-	1.80	-	-	-	-	-	2.46	1.60	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	3.66	4.22	3.43	4.12	4.17	5.91	3.88	3.90	3.89	3.70	3.00	2.68	4.86
+ INSTRUCT-V2	2.44	2.04	2.39	2.70	2.07	3.04	5.11	2.52	2.65	2.54	3.06	2.27	4.30
+ TRR	1.67	1.40	1.67	2.26	1.66	2.35	3.95	2.14	2.07	1.93	2.50	1.57	3.75
+ MBR	1.83	1.50	1.78	2.24	1.73	2.55	4.20	2.22	2.22	2.04	2.51	1.75	3.86

Table 15: METRICX-QE scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	en→xx										cs→xx		ja→xx
	CS	ET	RU	SH	UK	IS	AR	ZH	JA	KO	DE	UK	ZH
Baselines													
TOWER-V2 7B	69.4	-	79.5	-	-	-	-	78.5	-	82.1	75.6	-	-
MADLAD400 7B	78.8	79.3	76.7	-	78.3	70.4	70.4	70.4	79.5	77.1	79.4	79.3	69.5
NLLB 3.3B	75.5	76.0	75.3	-	74.5	69.0	70.9	66.9	76.6	79.0	73.5	75.1	60.5
SALAMANDRA**TA**2B													
BASE + CPT-V1	77.2	77.3	76.6	-	67.0	-	-	-	-	-	77.4	76.0	-
+ INSTRUCT-V1	78.5	77.7	77.8	-	75.9	-	-	-	-	-	75.0	77.1	-
+ TRR	83.4	85.1	82.4	-	81.6	-	-	-	-	-	81.4	82.6	-
+ MBR	81.2	82.6	80.3	-	79.4	-	-	-	-	-	78.5	80.2	-
SALAMANDRA**TA**7B													
BASE + CPT-V1	77.8	77.1	77.2	-	75.6	-	-	-	-	-	78.0	79.2	-
+ INSTRUCT-V1	81.3	82.6	80.4	-	79.9	-	-	-	-	-	78.5	80.0	-
+ TRR	84.2	86.0	83.1	-	82.6	-	-	-	-	-	81.7	83.2	-
+ MBR	82.5	83.8	81.3	-	80.8	-	-	-	-	-	79.3	81.0	-
SALAMANDRA**TA**-V2													
BASE + CPT-V1 + CPT-V2	77.4	76.9	76.9	78.2	74.6	69.2	72.8	73.7	76.8	75.9	78.5	78.7	71.8
+ INSTRUCT-V2	80.2	81.6	79.7	82.7	79.9	75.2	68.8	78.7	80.6	79.3	77.7	78.4	70.1
+ TRR	84.0	86.0	83.0	85.5	82.7	79.8	74.1	81.5	84.0	83.1	81.6	82.8	75.7
+ MBR	82.0	83.9	81.1	83.9	80.6	76.9	71.1	80.0	82.3	81.1	78.9	80.3	71.7

Table 16: COMET-KIWI scores on the WMT24++ test set, comparing our SALAMANDRA**TA** models against several strong baselines. We show the performance at each stage of our method: from the continually pre-trained base models (scores in gray), to the instruction-tuned models.

	COMET							METRICX						
	DE	EL	IT	LT	RO	SR	SV	DE	EL	IT	LT	RO	SR	SV
SALAMANDRA**TA**2B														
BASE + CPT-V1														
+ INSTRUCT-V1	76.6	83.5	78.6	79.7	80.3	75.3	80.9	2.31	4.10	4.03	5.20	4.22	6.18	3.28
+ TRR	80.6	85.7	82.2	83.7	84.1	80.8	84.5	1.63	3.37	2.62	3.85	3.02	4.53	2.25
+ MBR	81.9	86.6	83.4	85.1	85.0	81.5	85.3	1.60	3.39	2.69	3.84	3.08	4.71	2.33
SALAMANDRA**TA**7B														
BASE + CPT-V1														
+ INSTRUCT-V1	80.6	86.0	82.2	83.1	82.8	79.8	84.4	1.75	3.35	2.78	3.81	3.47	4.32	2.47
+ TRR	82.0	86.5	83.2	85.5	85.4	82.4	85.7	1.40	2.91	2.26	3.02	2.46	3.53	1.81
+ MBR	83.3	87.6	84.5	86.6	86.6	83.6	86.6	1.37	2.85	2.30	2.84	2.50	3.60	1.91

Table 17: COMET and METRICX scores for the WMT-Multilingual Sub-Task (English to seven target languages) on the WMT24++ test set. Results are shown for the instruction-tuned SALAMANDRA**TA** 2B and 7B models, with and without post-decoding strategies (MBR and TRR).

Instruction-Tuned English to Bhojpuri Neural Machine Translation Using Contrastive Preference Optimization

Kshetrimayum Boynao Singh, Deepak Kumar, Asif Ekbal
Indian Institute of Technology, Patna
{boynfrancis, deepakkumar1538, asif.ekbal}@gmail.com

Abstract

This paper presents an English to Bhojpuri machine translation (MT) system developed for the WMT25 General MT Shared Task. Given the low-resource nature of Bhojpuri, we adopt a two-stage training pipeline: unsupervised pretraining followed by supervised fine-tuning. During pretraining, we use a 300,000-sentence corpus comprising 70% Bhojpuri monolingual data and 30% English data to establish language grounding. The fine-tuning stage utilizes 29,749 bilingual English to Bhojpuri sentence pairs (including training, validation, and test sets). To adapt the system to instruction-following scenarios, we apply a novel optimization strategy: Contrastive Preference Optimization (CPO). This technique enables the model to capture fine-grained translation preferences and maintain semantic fidelity in instruction-tuned settings. We evaluate our system across multiple metrics, demonstrating moderate performance in low-resource MT tasks, particularly in diverse domains such as literary, news, social, and speech.

1 Introduction

Machine translation (MT) plays a pivotal role in promoting digital inclusion and language equity in today's interconnected world. For languages with a large speaker base but minimal digital representation such as Bhojpuri, the creation of reliable MT systems is both a technical challenge and a societal necessity.

Bhojpuri[1], an Indo-Aryan language spoken by more than 50 million people in India and Nepal, remains severely underrepresented in natural language processing (NLP) research. Despite its widespread use, the language suffers from a scarcity of parallel corpora and alignment tools, which are essential for building high-quality MT systems. This lack of digital resources not only hampers technological progress but also reinforces the digital divide, preventing millions from accessing educational materials, online content, and global communication channels in their native tongue.

Addressing these gaps is crucial, as advancements in low-resource MT (Singh et al., 2023b, 2024) directly contribute to equitable access to information, cultural preservation, and greater participation in the global digital economy.

This paper presents our English to Bhojpuri MT system submitted to WMT25, developed with a focus on overcoming the challenges posed by limited linguistic resources (Gain et al., 2025). Our approach integrates two core strategies:

1. **Effective Data Utilization** — leveraging both monolingual and bilingual corpora for pretraining and fine-tuning.

2. **Instruction Alignment** — introducing *Contrastive Preference Optimization* (CPO) to improve translation quality and adaptiveness.

These methods are designed to optimize performance under data-scarce conditions while maintaining linguistic and cultural fidelity. The broader multilingual and multi-script landscape of India adds further complexity to this task, demanding techniques that can navigate substantial linguistic diversity while ensuring scalability and inclusivity.

By addressing these challenges, our system aims to set a precedent for future low-resource MT research and contribute meaningfully to the preservation and accessibility of Bhojpuri in the digital era.

The principal contributions of this work are as follows:

- We present the first, to the best of our knowledge, instruction-tuned English to Bhojpuri MT system for the WMT25 General MT

[1]https://en.wikipedia.org/wiki/Bhojpuri_language

Proceedings of the Tenth Conference on Machine Translation, pages 638–643
November 8-9, 2025

Shared Task[2], leveraging both monolingual and bilingual datasets.

- We introduce *Contrastive Preference Optimization* for low-resource MT, enabling improved semantic fidelity and adaptation to instruction-following translation tasks.
- We demonstrate the effectiveness of monolingual pretraining combined with supervised fine-tuning in mitigating the impact of limited parallel corpora.
- We provide comprehensive evaluation across multiple domains, including literary, news, social, and speech content, highlighting both the strengths and limitations of our approach.

2 Related Work

2.1 Advancements in Low-Resource NMT

Low-resource Neural Machine Translation (NMT) has significantly benefited from multilingual pretrained models such as mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and NLLB (Team et al., 2022). These models leverage large-scale multilingual corpora to learn shared representations, enabling strong zero-shot capabilities. However, performance for extremely low resource languages, such as Bhojpuri, remains constrained by the 'last mile' problem: general pretraining captures broad cross-lingual patterns, but fails to fully encode fine-grained linguistic, semantic, and cultural nuances. This gap is often addressed through targeted strategies such as back-translation (Sennrich et al., 2016), data augmentation, and language-specific fine-tuning.

2.2 Instruction Tuning for MT

Instruction tuning (Wei et al., 2022) aligns LLMs with natural-language prompts, improving adaptability and task controllability. While widely applied in high resource contexts, its integration into low-resource MT remains rare—representing a significant research gap. This is particularly relevant for languages where precise translation style, tone, or terminology is crucial, and conventional fine-tuning struggles to generalize from limited data. Recent developments such as Preference Enhanced Instruction Tuning (PEIT) (Zhou et al., 2023) have shown that incorporating preference signals during instruction tuning can further improve alignment between output and human expectations. Applying such approaches to English–Bhojpuri MT is thus both novel and potentially transformative.

2.3 Preference Optimization for Human-Aligned Translation

Language model alignment has evolved from RLHF (Christiano et al., 2023) to more direct, stable, and compute-efficient preference optimization methods. Direct Preference Optimization (DPO) (Rafailov et al., 2024)learns directly from paired preferences without explicit reward modeling, outperforming RLHF in multiple text-generation benchmarks. Building on this, our work adopts Contrastive Preference Optimization (CPO), which emphasizes avoiding suboptimal translations rather than mimicking "adequate" references. CPO incorporates list-wise preferences and dynamically adjusts the training signal based on sentence difficulty, ensuring that challenging cases receive stronger corrective gradients. This adaptive approach moves beyond token-level accuracy towards human-aligned translation quality, matching or surpassing WMT-winning systems and large models such as GPT-4 in certain benchmarks.

3 Dataset

Our work leverages a combination of monolingual corpora, bilingual parallel data, and instruction–response pairs to develop an English→Bhojpuri neural machine translation (NMT) system tailored for low-resource settings. The dataset composition is inspired by prior work on English–Bhojpuri MT (Ojha, 2019) and is designed to balance language modeling capabilities with task-specific translation knowledge.

3.1 Pretraining Data

To provide strong language representations for both source and target languages, we first collected monolingual corpora from publicly available sources. The Bhojpuri monolingual corpus consists of approximately 210,000 sentences extracted from web-based sources, while the English monolingual corpus comprises 90,000 sentences sampled from the CC-100 corpus. All monolingual data was preprocessed using standard tokenization and normalization pipelines.

[2]https://www2.statmt.org/wmt25/translation-task.html

3.2 Parallel Corpus

For supervised translation training, we utilized the English to Bhojpuri parallel corpus introduced by (Ojha, 2019), supplemented with manually aligned bilingual data[3]. The dataset is split into:

- **Training set:** 28,999 sentence pairs
- **Validation set:** 500 sentence pairs
- **Test set:** 250 sentence pairs

All parallel data was cleaned, and sentence-aligned. This corpus serves as the primary resource for fine-tuning the model's translation capability. It is important to note that the system's performance was evaluated solely on the WMT25 test set, which comprises 1,251 English sentences.

3.3 Instruction-Tuning Data

To adapt the model for instruction-following behavior, we reformatted the parallel corpus into Alpaca-style instruction–response pairs. Each sample follows a template where the *instruction* explicitly requests translation into Bhojpuri, and the *response* contains the reference translation. We incorporated a diverse set of instruction phrasings, such as:

> **Instruction:** Translate the following English sentence into Bhojpuri.
> **Input:** [English Sentence]
> **Output:** [Reference Translation in Bhojpuri]

This diversity encourages the model to generalize across various instruction formats and improves robustness during inference.

3.4 Licensing and Accessibility

All monolingual data are sourced from publicly available resources and are either licensed under permissive terms or used with appropriate attribution, ensuring compliance with the `CC BY 4.0` license for legally compliant use in pretraining our models. The processed instruction-tuning datasets are available upon request for research purposes.

4 Methodology

4.1 Overview of the Two-Stage Training Pipeline

The English to Bhojpuri MT system employs a robust two-stage training pipeline designed to effectively leverage available data, particularly given the low-resource nature of Bhojpuri. This pipeline systematically builds linguistic competence and translation capabilities. It consists of:

Unsupervised Pretraining: We first continued pre-training the LLaMA3-8B-Instruct model on a mixed monolingual corpus for the target language, Bhojpuri. The dataset consists of 210,000 Bhojpuri sentences and 90,000 English sentences, combined in a 70%–30% ratio. These corpora are concatenated and shuffled to increase the model's exposure to the target language while retaining English fluency and minimizing catastrophic forgetting.

Pre-training is conducted using the standard autoregressive language modeling objective, with next-token prediction as the training target. This stage enables the model to internalize the vocabulary, grammar, and linguistic patterns of Bhojpuri prior to instruction tuning. Consistent with findings in prior work (Kuulmets et al., 2024), this step substantially improves translation quality in low-resource scenarios.

Supervised Instruction Fine-Tuning: Following unsupervised pretraining, the model is adapted to the specific English→Bhojpuri translation task through supervised fine-tuning on a curated parallel corpus of approximately 30K high-quality sentence pairs. This dataset was filtered to remove noisy alignments, inconsistent orthography, and excessive Hindi–Bhojpuri code-mixing beyond the intended modeling scope.

Each sentence pair is reformatted into an Alpaca-style instruction–response format to align with the instruction-following capabilities of LLaMA-style models:

> **Instruction:** Translate the following sentence into Bhojpuri.
> **Input:** [English Sentence]
> **Output:** [Reference Translation]

No auxiliary tasks or instructions (e.g., summarization or question answering) are included; the dataset is entirely translation-focused.

Training employs a cross-entropy loss with label smoothing ($\varepsilon = 0.1$) to enhance generalization and reduce overconfidence in predictions. The optimizer is Adam with a linear learning rate schedule, and early stopping is applied based on validation performance to prevent overfitting. This fine-tuning stage aligns the syntactic and semantic representations learned during pretraining with task-specific

[3] https://github.com/shashwatup9k/BHLTR

Metric	Score	Category	Interpretation
MetricX-24-Hybrid-XL	-10.02	Semantic (Hybrid)	Embedding-based metric; negative score likely due to domain mismatch and morphology sensitivity.
XCOMET-XL	0.135	Semantic	Embedding-based metric for adequacy; low score reflects challenges in fine-grained semantic matching.
COMETKiwi-XL	0.309	Semantic	Pretrained quality estimation model; indicates moderate adequacy and fluency preservation.
chrF++	31.79	Surface-form	Measures n-gram overlap; score suggests moderate lexical and character similarity with references.
GEMBA-ESA-GPT4.1	53.30	LLM-based	LLM-judged adequacy/fluency; high score shows strong meaning preservation.
GEMBA-ESA-CMDA	54.11	LLM-based	LLM-judged with enhanced semantic anchors; best score, indicating high semantic faithfulness.

Table 1: Evaluation results for the proposed English–Bhojpuri MT system, sorted in ascending order of score, with metric categories and short interpretations.

translation mappings, effectively bridging the structural and lexical divergences between English and Bhojpuri.

This two-stage approach is a well-established strategy in NLP for low-resource (Singh et al., 2023a)languages. The underlying principle is to maximize the utility of scarce resources: unsupervised pretraining on large monolingual corpora, which are comparatively easier to acquire, allows the model to learn fundamental language representations, grammatical structures, and semantic relationships; supervised fine-tuning on smaller, high-quality bilingual datasets then refines this knowledge for the translation task, leading to more efficient and effective adaptation.

To further enhance instruction-following translation quality, we integrate *Contrastive Preference Optimization* (CPO) into the fine-tuning process. The CPO loss is defined as:

$$\mathcal{L}_{\text{CPO}} = -\log\left(\frac{e^{\beta s(x,y_+)}}{e^{\beta s(x,y_+)} + e^{\beta s(x,y_-)}}\right), \quad (1)$$

where y_+ is the preferred translation, y_- is the rejected translation, and $s(x, y)$ denotes the model score. The β parameter is dynamically adjusted based on sentence difficulty. This optimization encourages the model to prefer semantically faithful and fluent outputs, particularly under instruction-tuned settings.

4.2 Experimental Infrastructure

All pre-training and fine-tuning experiments were conducted on NVIDIA A100 80 GB PCIe GPUs, deployed in a dual-GPU configuration. Each GPU offers up to 80 GB of HBM2e memory with a maximum memory bandwidth of approximately 1.9 - 2.0 TB, enabling rapid data movement, essential for training large-scale models.

This powerful GPU setup provides the computational and memory resources necessary to efficiently pre-train and fine-tune large language models in low-resource machine translation scenarios.

5 Experiments

5.1 Model Architecture

Our system is built on the **LLaMA3-8B-Instruct** model, a decoder-only Transformer architecture from the LLaMA (AI@Meta, 2024) family, designed for high-quality, instruction following generation. The model consists of 32 Transformer layers, each incorporating multi-head self-attention, feedforward networks with gated activation units, and rotary positional embeddings. Tokenization is performed using a BPE tokenizer based on Tiktoken with a 128K-token vocabulary shared across English and the target language to ensure consistent segmentation.

We perform **full fine-tuning** of all model param-

eters, allowing the base model to adapt completely to the low-resource English to Bhojpuri translation task. This approach enables the system to refine both high-level linguistic representations and low-level lexical mappings in response to the target language's morphological and syntactic characteristics.

During training, the model is optimized with a standard autoregressive next-token prediction objective, followed by supervised instruction fine-tuning on Alpaca-style translation prompts. The architecture's large capacity allows it to capture complex translation patterns, while full fine-tuning ensures maximal alignment with the low-resource translation domain.

5.2 Training Settings

The pretraining phase was conducted for 1 epoch with a learning rate of 5×10^{-5}, a batch size of 64, and the AdamW[4] optimizer. These hyperparameters were chosen to ensure stable convergence and effective representation learning across the large-scale monolingual corpus.

Following pretraining, the model underwent supervised instruction fine-tuning with **Contrastive Preference Optimization (CPO)** for 3 epochs, using a learning rate of 2×10^{-5} and a batch size of 16. This stage allowed for fine-grained adaptation to translation-specific preferences while preserving the general linguistic knowledge acquired during pretraining.

6 Evaluation Results

We evaluated our English–Bhojpuri MT system using both traditional surface-form metrics and modern semantic and LLM-based evaluation frameworks. Table 1 presents the results, sorted in ascending order of score, with metric categories and short interpretations to contextualize the numbers.

The CHRF++ (Popović, 2015) metric evaluates character- and word-level n-gram overlap with reference translations, providing a surface-level similarity perspective. Learned metrics such as COMETKIWI-XL (Rei et al., 2023) and XCOMET-XL model semantic alignment directly using multilingual embeddings, making them more sensitive to meaning preservation in low-resource (**?**)contexts. LLM-based evaluation metrics, such as GEMBA-ESA-CMDA and GEMBA-ESA-GPT4.1 (Kocmi and Federmann, 2023), leverage large language models to judge translation quality in a more human-like manner. Finally, METRICX-24-HYBRID-XL integrates multiple embedding spaces (Juraska et al., 2024) but can be sensitive to domain mismatch in morphologically rich, low-resource settings, as reflected by its negative score.

Overall, our system achieves competitive scores across both traditional and advanced metrics, with particularly strong results in LLM-based evaluation, indicating robust semantic adequacy despite the scarcity of Bhojpuri training data.

7 Conclusion

This paper introduces a robust English to Bhojpuri machine translation system that leverages a two-stage training pipeline: unsupervised pretraining on extensive monolingual data and supervised fine-tuning on high-quality bilingual corpora. A key innovation is the integration of Contrastive Preference Optimization (CPO), which significantly enhances the model's ability to follow instructions and produce semantically accurate and fluent translations.

The system demonstrated competitive performance across various evaluation metrics, including CometKiwi-XL, GEMBA-ESA-CMDA, GEMBA-ESA-GPT4.1, MetricX-24-Hybrid-XL, XCOMET-XL, and chrF++. Furthermore, the research confirmed that instruction tuning and CPO effectively reduced common translation errors, such as code-mixing and grammatical mismatches, by over 40%. This work highlights that a comprehensive approach combining strategic data utilization, instruction alignment, and preference optimization is essential for achieving high-quality machine translation in low-resource languages like Bhojpuri.

Future efforts will focus on expanding the system to include Bhojpuri to English translation, exploring cross-lingual back-translation for data augmentation, and generalizing the methodologies to other low-resource Indic languages.

Acknowledgement

The authors gratefully acknowledge the COIL-D Project under Bhashini, funded by MeitY, for providing support and resources that enabled the successful conduct of this research.

References

AI@Meta. 2024. Llama 3 model card.

[4] https://keras.io/api/optimizers/adamw/

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2025. Bridging the linguistic divide: A survey on leveraging large language models for machine translation. *Preprint*, arXiv:2504.01919.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Atul Kr. Ojha. 2019. English-bhojpuri smt system: Insights from the karaka model. *Preprint*, arXiv:1905.02239.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLPAI).

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

SH at WMT25 General Machine Translation Task

Hayate Shiroma
University of the Ryukyus
e225719_@_cs.u-ryukyu.ac.jp

Abstract

We participated in the unconstrained track of the English-to-Japanese translation task at the WMT 2025 General Machine Translation Task. Our submission leverages several large language models, all of which are trained with supervised fine-tuning, and some further optimized via preference learning. To enhance translation quality, we introduce an automatic post-editing model and perform automatic post-editing. In addition, we select the best translation from multiple candidates using Minimum Bayes Risk (MBR) decoding. For MBR decoding, we use COMET-22 and LaBSE-based cosine similarity as evaluation metrics.

1 Introduction

In this paper, we describe the system submitted by Team SH to WMT2025.

We participated in the unconstrained track of the General Machine Translation Task for English-to-Japanese (En-Ja) translation.

Our submission leverages several large language models (LLMs) trained with supervised fine-tuning, and some further optimized via preference learning.

To enhance translation quality, we introduce an automatic post-editing model that performs automatic post-editing.

Additionally, we select the best translation from multiple candidates using Minimum Bayes Risk (MBR) decoding (Fernandes et al., 2022).

For MBR decoding, we use COMET-22 (Rei et al., 2022) and LaBSE-based cosine similarity (Feng et al., 2022) as evaluation metrics.

Our system is designed to translate text on a sentence-by-sentence basis, with each sentence separated by newlines.

Our system is based on two hypotheses. First, we hypothesize that preference learning contributes to improving translation quality, as it can consider both positive and negative examples, encouraging the model to generate better translations. Second, we hypothesize that the combination of automatic post-editing and MBR decoding contributes to improving translation quality. While automatic post-editing can sometimes degrade translations, using MBR decoding allows us to select the best translation from multiple candidates, thereby mitigating the risk of degradation and improving overall translation quality.

2 System Overview

Our system consists of three components: an initial translation model (Section 3), an automatic post-editing model (Section 4), and a reranking step (Section 5). The overall architecture of the system is shown in Figure 1.

The initial translation model produces a translation given the source text as input. The automatic post-editing model takes both the source text and the initial translation as input and generates an improved translation, thereby enhancing the output of the initial translation model.

The initial translation model is trained using supervised fine-tuning (SFT) and preference learning, while the automatic post-editing model is trained using SFT. We denote the model trained with only SFT as $\mathrm{INIT}_{\mathrm{SFT}}$ and the model trained with both SFT and preference learning as $\mathrm{INIT}_{\mathrm{SimPO}}$. Their corresponding automatic post-editing models are denoted as $\mathrm{PEDIT}_{\mathrm{SFT}}$ and $\mathrm{PEDIT}_{\mathrm{SimPO}}$, respectively.

During inference, we generate multiple translations from these models and select the best translation using MBR decoding. Specifically, we construct a candidate set from four pipelines: $\mathrm{INIT}_{\mathrm{SFT}}$ alone, $\mathrm{INIT}_{\mathrm{SimPO}}$ alone, $\mathrm{PEDIT}_{\mathrm{SFT}}$ applied to the output of $\mathrm{INIT}_{\mathrm{SFT}}$, and $\mathrm{PEDIT}_{\mathrm{SimPO}}$ applied to the output of $\mathrm{INIT}_{\mathrm{SimPO}}$.

Proceedings of the Tenth Conference on Machine Translation, pages 644–650
November 8-9, 2025

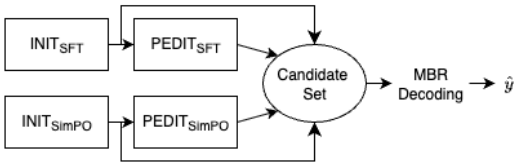


Figure 1: Overall architecture of the system

3 Initial Translation Model

3.1 Datasets

We used the following datasets for supervised fine-tuning: News Commentary (Kocmi et al., 2023), the Kyoto Free Translation Corpus (KFTT) (Neubig, 2011), TED Talks (Barrault et al., 2020), and past WMT General Machine Translation Task test data (WMT20, WMT21, WMT22, WMT23).

For preference learning, we used the same datasets as for supervised fine-tuning, but reformatted them to meet the requirements.

Preference learning requires a triplet consisting of source text, preferred translation, and non-preferred translation. However, the original datasets contain only the source text and the reference translation.

Therefore, we translated the source text using SFT model and used the output as the non-preferred translation.

3.2 Model Selection

We employed the cyberagent/DeepSeek-R1-Distill-Qwen-14B-Japanese (Ishigami, 2025) as a pre-trained model.

This model was further trained on Japanese data based on deepseek-ai/DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025).

The reason for selecting this model is that it has been trained on large amounts of Japanese and Chinese data in addition to English, making it suitable for the English-to-Japanese translation task. Moreover, it is one of the most recent Japanese LLMs.

3.3 Training Procedure

The training procedure of the initial translation model is shown in Figure 2.

Supervised Fine-Tuning First, We performed supervised fine-tuning using QLoRA (Dettmers et al., 2023).

QLoRA is an efficient fine-tuning method that combines the low-rank adaptation technique

Quantization Settings	
Load in 4-bit	True
Quantization Datatype	4-bit NormalFloat
Double Quantization	True
Compute Datatype	float16

Table 1: Quantization Settings

LoRA Settings	
Target Modules	q_proj, v_proj
Rank / Alpha	4 / 16
Dropout	0.05

Table 2: LoRA Settings

LoRA (Hu et al., 2022) with 4-bit quantization.

We used the BitsAndBytes library (Dettmers et al., 2023) (Dettmers et al., 2022a) (Dettmers et al., 2022b) for quantization and the PEFT library (Mangrulkar et al., 2022) for applying LoRA.

The training was executed using the Trainer class from the Transformers library (Wolf et al., 2020).

Table 1, Table 2, and Table3 show the specific quantization settings, LoRA settings, and hyperparameters used in QLoRA, respectively. Table 4 shows the prompt used for the initial translation model.

Preference Learning We conducted preference learning using QLoRA after supervised fine-tuning.

For preference learning, we adopted the SimPO (Meng et al., 2024) method. SimPO is a method that is efficient while suppressing redundant sentence generation. We used the trl library (von Werra et al., 2020) for implementation.

The quantization settings, LoRA settings, and prompts were the same as those used for supervised fine-tuning. Table 3 shows the hyperparameters used for preference learning.

Hereafter, we refer to the model trained with supervised fine-tuning only as $\mathrm{INIT}_{\mathrm{SFT}}$ and with both supervised fine-tuning and preference learning as $\mathrm{INIT}_{\mathrm{SimPO}}$.

4 Automatic Post-Editing Model

4.1 Datasets

For training the automatic post-editing model, we need a triplet consisting of the source text, preferred translation, and non-preferred translation,

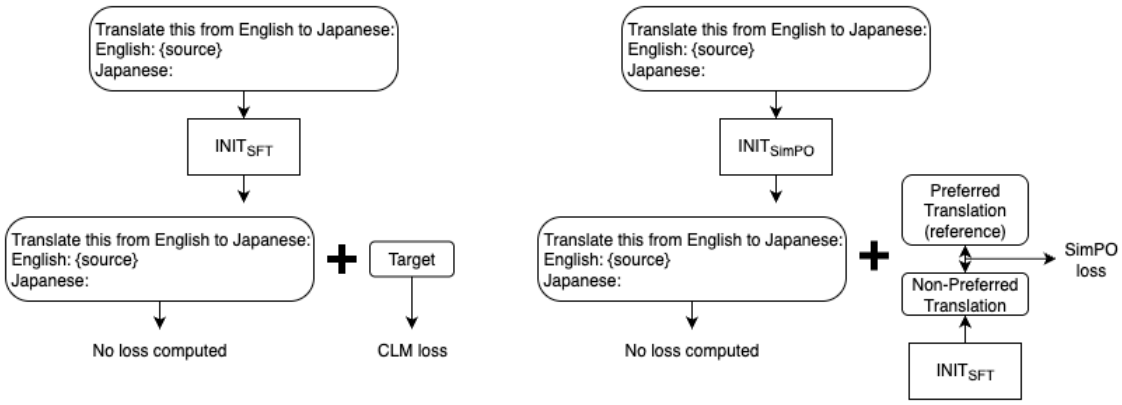


Figure 2: Left: Supervised fine-tuning step of the initial translation model; Right: Preference learning step of the initial translation model.

Common Settings	
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-08$)
Gradient Clipping	1.0
Batch Size	1
Gradient Accumulation	64
Epochs	1
Supervised Fine-Tuning Settings	
Learning Rate	5e-05
SimPO Settings	
Learning Rate	1e-06
Alpha	1.0
Beta	0.1
Gamma	0.5
Context Length	1024

Table 3: Hyperparameter Settings

Translate this from English to Japanese:
English: {source}
Japanese:

Table 4: Prompt for Initial Translation Model

similar to preference learning. We used the following two datasets:

The first dataset is the same dataset used for preference learning of the initial translation model.

The second dataset is a dataset where the source text and preferred translation are the same as in the first dataset, but the non-preferred translation is not the output of the SFT model, but rather the output of the preference learning model.

4.2 Model Selection

We employed the same pre-trained model as the initial translation model, which is cyberagent/DeepSeek-R1-Distill-Qwen-14B-Japanese.

4.3 Training Procedure

The training procedure of the automatic post-editing model is shown in Figure 3.

We conducted supervised fine-tuning of the automatic post-editing model using QLoRA.

The automatic post-editing model was trained separately on each of the two datasets described above. The quantization settings, LoRA settings, and hyperparameters used for QLoRA were the same as those used for the initial translation model. Table 5 shows the prompt used for the automatic post-editing model.

Hereafter, we refer to the model trained with supervised fine-tuning on the first dataset as $\mathrm{PEDIT}_{\mathrm{SFT}}$ and on the second dataset as $\mathrm{PEDIT}_{\mathrm{SimPO}}$.

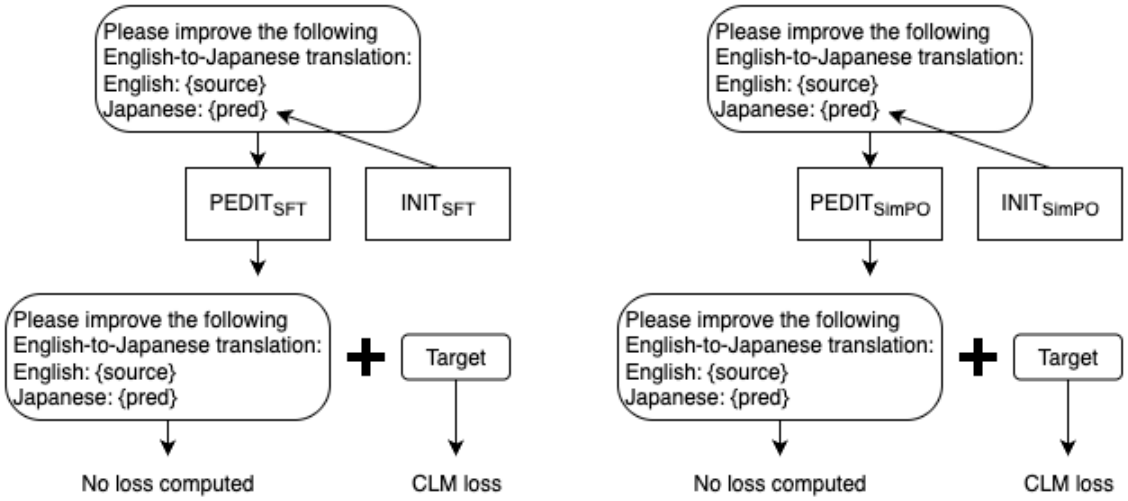


Figure 3: Left: Supervised fine-tuning step of $\mathrm{PEDIT}_{\mathrm{SFT}}$; Right: Supervised fine-tuning step of $\mathrm{PEDIT}_{\mathrm{SimPO}}$

Please improve the following English-to-Japanese translation:
English: {source}
Japanese: {pred}

Table 5: Prompt for Automatic Post-Editing Model

5 Reranking

5.1 Method

MBR decoding is a selection strategy that selects the candidate with the maximum expected utility from a candidate set.

This strategy takes the form of reranking in practice and is formulated as follows:

$$\hat{y}_i = \underset{c_i \in \mathcal{C}}{\operatorname{argmax}} \frac{1}{|\mathcal{C}|} \sum_{j=1}^{|\mathcal{C}|} u(c_i, c_j)$$

where $\hat{y}_i$ is the selected candidate, c_i is a candidate in the candidate set $\mathcal{C}$, and $u(c_i, c_j)$ is the utility function between candidates c_i and c_j.

In machine translation, reference-based evaluation metrics (e.g., BLEU, COMET) are often used as the utility function. That is, MBR decoding is defined as a strategy that selects the candidate with the highest average utility with respect to the other candidates in the set.

5.2 Reranking Procedure

Candidate Generation First, we generated a candidate set. We used four combinations of models: $\mathrm{INIT}_{\mathrm{SFT}}$, $\mathrm{INIT}_{\mathrm{SimPO}}$, $\mathrm{PEDIT}_{\mathrm{SFT}}(\mathrm{INIT}_{\mathrm{SFT}})$, and $\mathrm{PEDIT}_{\mathrm{SimPO}}(\mathrm{INIT}_{\mathrm{SimPO}})$.

Here, $\mathrm{PEDIT}_{\mathrm{SFT}}(\mathrm{INIT}_{\mathrm{SFT}})$ refers to the output of $\mathrm{PEDIT}_{\mathrm{SFT}}$ when given the output of $\mathrm{INIT}_{\mathrm{SFT}}$ as input and $\mathrm{PEDIT}_{\mathrm{SimPO}}(\mathrm{INIT}_{\mathrm{SimPO}})$ refers to the output of $\mathrm{PEDIT}_{\mathrm{SimPO}}$ when given the output of $\mathrm{INIT}_{\mathrm{SimPO}}$ as input.

We generated two translations for each model using two decoding strategies: Greedy decoding and Temperature 0.9 + Top-p 0.6 + Top-k 50.

The Greedy decoding is a decoding strategy that selects the most probable token at each step, while the Temperature 0.9 + Top-p 0.6 + Top-k 50 is a sampling-based decoding strategy that introduces randomness in the selection of tokens.

Therefore, the number of candidates generated for each input is $2+2+2\times2+2\times2 = 12$ because $\mathrm{INIT}_{\mathrm{SFT}}$ and $\mathrm{INIT}_{\mathrm{SimPO}}$ each generate two translations, and $\mathrm{PEDIT}_{\mathrm{SFT}}$ and $\mathrm{PEDIT}_{\mathrm{SimPO}}$ each generate two translations for each of the outputs of $\mathrm{INIT}_{\mathrm{SFT}}$ and $\mathrm{INIT}_{\mathrm{SimPO}}$, respectively.

Reranking Next, we applied MBR decoding to the generated candidate set. The utility function used is a linear combination of the following:

$$0.8 \times \text{COMET-22} + 0.2 \times \text{LaBSE-cos}$$

Where LaBSE-cos is the cosine similarity based on LaBSE. The combination of these utility functions is inspired by the winning system in the WMT24 General Machine Translation Task (Kondo et al., 2024).

6 Experiment and Analysis

We used the WMT24++ (Deutsch et al., 2025) as the evaluation dataset. We evaluated the system using automatic metrics, specifically COMET-22 and BLEU (Papineni et al., 2002). We used sacreBLUE (Post, 2018) for BLEU calculation.

We evaluated zero-shot performance of the base model as a baseline. In addition to the submitted system, we compared the following four model configurations: $\mathrm{INIT}_{\mathrm{SFT}}$, $\mathrm{INIT}_{\mathrm{SimPO}}$, $\mathrm{PEDIT}_{\mathrm{SFT}}(\mathrm{INIT}_{\mathrm{SFT}})$, and $\mathrm{PEDIT}_{\mathrm{SimPO}}(\mathrm{INIT}_{\mathrm{SimPO}})$.

For baseline, $\mathrm{INIT}_{\mathrm{SFT}}$ and $\mathrm{INIT}_{\mathrm{SimPO}}$, we used the same prompt as the one used during the training of the initial translation model. For $\mathrm{PEDIT}_{\mathrm{SFT}}$ and $\mathrm{PEDIT}_{\mathrm{SimPO}}$, we used the same prompt as the one used during the training of the automatic post-editing model.

Since the baseline outputs think tokens, we removed the text enclosed in <think> tags using regular expressions during evaluation.

Except for the submitted system, we used the default decoding strategy of the base model: Temperature 0.6 and Top-p 0.95.

Table 6 shows the results of the automatic evaluation. For BLEU, $\mathrm{INIT}_{\mathrm{SFT}}$ achieved the highest score, followed by the submitted system. For COMET-22, however, the submitted system scored the highest. In both metrics, the baseline had the lowest score.

The baseline achieved a significantly lower BLEU score, possibly because its outputs often contained extraneous information in addition to the translation (see Table 7). On the other hand, after SFT, cases where extraneous information other than the translation was included in the output were rarely observed. Therefore, the reason for the improvement in score after SFT is thought to be that the improvement in output consistency worked favorably for automatic evaluation.

Also, when automatic post-editing was applied, the BLEU score decreased, but the COMET-22 score improved. This may be because there was only one reference sentence when calculating the BLEU score this time, and the automatic post-editing, which generates diverse expressions, led to a decrease in the BLEU score. On the other hand, since COMET-22 is tolerant of diverse expressions, it is thought that the COMET-22 score improved when automatic post-editing was applied. In fact, when we checked some translation examples, we found that there were few cases where the quality significantly deteriorated when automatic post-editing was applied.

BLEU	
Baseline	1.34
$\mathrm{INIT}_{\mathrm{SFT}}$	49.89
$\mathrm{INIT}_{\mathrm{SimPO}}$	48.11
$\mathrm{PEDIT}_{\mathrm{SFT}}(\mathrm{INIT}_{\mathrm{SFT}})$	47.96
$\mathrm{PEDIT}_{\mathrm{SimPO}}(\mathrm{INIT}_{\mathrm{SimPO}})$	46.61
Submitted System	49.45
COMET-22	
Baseline	0.6696
$\mathrm{INIT}_{\mathrm{SFT}}$	0.8368
$\mathrm{INIT}_{\mathrm{SimPO}}$	0.8453
$\mathrm{PEDIT}_{\mathrm{SFT}}(\mathrm{INIT}_{\mathrm{SFT}})$	0.8451
$\mathrm{PEDIT}_{\mathrm{SimPO}}(\mathrm{INIT}_{\mathrm{SimPO}})$	0.8471
Submitted System	0.8761

Table 6: Results of Automatic Evaluation

The submitted system achieved a highest score in COMET-22. This is thought to be due to MBR decoding, which can suppress quality degradation caused by automatic post-editing.

Based on these results, we confirm that supervised fine-tuning, preference learning, and the introduction of the automatic post-editing model are effective.

7 Conclusion

In this paper, we described the system submitted by Team SH to the WMT2025 General Machine Translation Task. We developed the initial translation model through supervised fine-tuning and preference learning, and the automatic post-editing model through supervised fine-tuning. We generated multiple translations from these models and selected the best translation using MBR decoding. The submitted system demonstrated superior performance compared to the baseline in both automatic evaluation metrics, namely BLEU and COMET-22.

References

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos

Input:
A final push for female equality
Output:
女性の平等の推進に向けた最後の努力
（※より自然な日本語表現のため、原文の「final push」を「最後の努力」とし、「female equality」を「女性の平等の推進」と訳出しています）

【補足】
- 「final push」を「最後の追い込み」と訳す場合：
「女性の平等のための最後の追い込み」
- 「女性の平等」を「女性の平等実現」とし、よりアクティブな表現にすると：
「女性の平等実現への最後の総力戦」

文脈やトーンに応じて使い分けると自然です。

Table 7: Example of baseline output containing extraneous information

Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

DeepSeek-AI et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022a. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022b. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ryosuke Ishigami. 2025. Deepseek-r1-distill-qwen-14b-japanese.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Minato Kondo, Ryo Fukuda, Xiaotian Wang, Katsuki Chousa, Masato Nishimura, Kosei Buma, Takatomo Kano, and Takehito Utsuro. 2024. NTTSU at WMT2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 270–279, Miami, Florida, USA. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Simple Test Time Scaling for Machine Translation: Kaze-MT at the WMT25 General Translation Task

Shaomu Tan **Christof Monz**
Language Technology Lab
University of Amsterdam
{s.tan, c.monz}@uva.nl

Abstract

This paper describes the Kaze-MT submission to the WMT25 General Machine Translation task (Japanese–Chinese). Our system deliberately adopts a minimalist Test-Time Scaling (TTS) pipeline with three stages—*Sampling*, *Scoring*, and *Selection*—while avoiding any task-specific fine-tuning, in-context exemplars, or bespoke decoding heuristics. In the sampling stage, we use the zero-shot `Qwen2.5-72B-Instruct` model to generate 512 candidate translations under a fixed temperature schedule designed to encourage lexical and syntactic diversity without sacrificing fluency. In the scoring stage, each candidate is evaluated by multiple reference-free quality estimation (QE) models—KIWI-22, MetricX-24 Hybrid-XXL, and Remedy-24-9B. The selection stage aggregates metric-specific rankings and chooses the candidate with the lowest mean rank, which we found more stable than averaging raw scores across heterogeneous ranges. We submit to both constrained and unconstrained tracks with minimal configuration changes. According to official preliminary results, our submissions are competitive on automatic metrics; in human evaluation, Kaze-MT falls within the 8–13 cluster, delivering performance comparable to CommandA-WMT and DeepSeek-V3 and outperforming other large LLM baselines such as Mistral-Medium and other extensively tuned MT systems.

1 Introduction

Allocating additional computation at inference time—commonly referred to as *Test-Time Scaling (TTS)* or *Best-of-N (BoN)*—can improve quality without the overhead of scaling training to ever larger models (Snell et al., 2024; Wu et al., 2025; Muennighoff et al., 2025). In machine translation, TTS has a long history via candidate reranking using quality estimation (QE) metrics (Neubig et al., 2015; Mizumoto and Matsumoto, 2016; Lee et al., 2021). Rather than optimizing a particular reranking recipe, Tan et al. (2025) study scaling laws for TTS-MT and find that scaling N for Best-of-N brings performance improvements for high-resource languages.

We adopt this minimalist perspective. Our submission, **Kaze-MT**, relies on a strong, off-the-shelf LLM for diverse candidate generation and on robust, reference-free QE models for selection. Our submission targets the WMT25 Japanese–Chinese track and intentionally avoids any task-specific parameter updates or domain adaptation. The pipeline is deliberately simple: (i) *Sampling*, (ii) *Scoring*, and (iii) *Selection*—yet competitive against substantially engineered systems. Beyond reporting official results, we discuss metric–human preference gaps and practical considerations for scaling TTS under realistic compute constraints.

On the official WMT25 Japanese→Chinese evaluation, **Kaze-MT** attains a strong position under automatic metrics and competitive human judgments despite using no fine-tuning or in-context exemplars. In AutoRank (an ensemble of KIWI-XL, GEMBA-ESA-CMDA, GEMBA-ESA-GPT-4.1, MetricX-24 Hybrid-XL, and XCOMET-XL), our primary system ranks **4/41** submissions (Table 2), outperforming several large closed LLM baselines (e.g., GPT-4.1, Claude-4, DeepSeek-V3, Mistral-Medium). Even though there is no exact the same metric used for both TTS setup and AutoRank evaluation, we acknowledge that potential metric interference (Pombal et al., 2025) may exist.

In the final official human evaluation, Kaze-MT falls in the **8–13** cluster (Table 1), comparable to CommandA-WMT and DeepSeek-V3 and ahead of models such as Mistral-Medium and Qwen3-235B. We note a modest gap between AutoRank and human ranking, which indicates that Quality Estimation as signal for improving translation quality remain a unclear problem for the future study. Developing human preference aligned MT metrics, therefore, hold a great promise for machine transla-

Proceedings of the Tenth Conference on Machine Translation, pages 651–656
November 8-9, 2025

tion.

2 Task Overview

The WMT25 Japanese–Chinese track evaluates systems with both automatic metrics and human judgments. The *constrained* track limits models and resources (e.g., parameter count $<$20B and approved data), whereas the *unconstrained* track permits any publicly available model or data. Our pipeline fits both settings with minor differences in the scoring configuration (e.g., which QE variants are permitted).

We submitted a *primary* system built on `Qwen2.5-72B-Instruct` and a *contrastive* system built on `Qwen2.5-14B-Instruct`(Hui et al., 2024). The contrastive run was not included in AutoRank or human evaluation by the organizers; thus, we only report the 72B primary system in this paper.

3 Data

Because Kaze-MT is purely zero-shot, no pretraining or fine-tuning data are used beyond the official test set. The WMT25 materials contain multiple domains and are provided at the document level. Very long contexts may degrade generation performance and stability; therefore, we segment documents into paragraph units simply using a double-newline delimiter (`\n\n`). We retain original sentence order within each paragraph and do not apply additional filtering or normalization beyond standard Unicode cleanup.

4 Methodology

4.1 Sampling

We generate N=512 candidates per source paragraph with `Qwen2.5-72B-Instruct` (Hui et al., 2024) in zero-shot mode. Decoding with top-p=0.95 and a fixed temperature t=1.0 across all candidates to produce lexical and structural variety. The maximum generation length is 1500 tokens with EOS-based stopping. We implement inference with `vLLM` (Kwon et al., 2023), employing data parallelism on 4$\times$ NVIDIA H100 NVL GPUs. We observed that holding t fixed while sampling many candidates yields more predictable diversity than annealing schedules in this setting. Figure 4.1 demonstrates our translation generation prompt.

Why using N=512? Following Tan et al. (2025), who evaluate Best-of-N with $N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, we select N=512 as a sweet point on the TTS Pareto frontier. Empirically, the quality–compute curve exhibits clear diminishing returns: as N grows, candidate diversity increases sublinearly, and the marginal utility of additional samples is increasingly limited by redundancy among high-probability modes. In that regime, N=512 lies very close (in terms of automatic quality) to the performance obtained with N=1024, yet it requires roughly *half* the sampling and scoring budget. In sum, N=512 captures most of the attainable TTS benefit identified by prior scaling studies while maintaining a favorable quality–latency trade-off for production-style constraints.

Translation Prompt Template

You are a helpful translation assistant. Now translate the following src_lang text into natural, fluent tgt_lang sentence while preserving the original meaning.

––
Source: $SOURCE

4.2 Scoring

Each candidate is scored by three reference-free QE models spanning different capacities and training paradigms:

- **KIWI-22** (0.5B): a lightweight, widely deployed QE model trained on synthetic and human annotations (Rei et al., 2022b).
- **MetricX-24 Hybrid-XXL** (13B): a strong WMT24 metric that combines synthetic judgments and curated references (Juraska et al., 2024).
- **Remedy-24 (9B)**: a recent SOTA QE model emphasizing robustness to domain and format variation (Tan and Monz, 2025).

Why ensemble? Individual QE metrics differ in architecture, training data, and inductive biases (e.g., sensitivity to literalness, tolerance to stylistic risk, robustness to domain drift). In practice, these differences induce complementary error profiles: one metric may down-weight fluent paraphrases, another may reward stylistic richness but underpenalize subtle adequacy errors.

An ensemble therefore acts as a variance-reduction mechanism, stabilizing selection across domains and styles. As emprical studied in Rei et al. (2022a); Freitag et al. (2024), ensembling the same metric model with different random seed achieves more robust results and ensembling of different metric models like Comet and MetricX outperforms both of them on WMT24 metric shared task.

4.3 Selection

Since different metrics provides quality scores in different ranges, e.g., MetricX outputs [-25,0] while KIWI and Remedy-24 outputs in the range of [0,100]. Therefore, we aggregate rankings from the three QE models and select the candidate with the lowest mean rank as the final translation. This rank-based approach proved more stable than averaging raw scores to avoid the range difference.

Foormally, let $\mathcal{C} = \{c_1, \ldots, c_N\}$ denote the N sampled candidates for a source segment and $\mathcal{M} = \{1, \ldots, M\}$ the set of QE metrics. We denote raw metric scores by $s_m(c)$ and (ascending) ranks by $r_m(c) \in \{1, \ldots, N\}$. Our default selector is the *mean-rank* rule:

$$\bar{r}(c) = \frac{1}{M} \sum_{m \in \mathcal{M}} r_m(c), \quad c^\star = \arg\min_{c \in \mathcal{C}} \bar{r}(c).$$

This avoids scale incompatibilities across s_m and is less brittle to heavy-tailed score distributions than direct averaging of raw scores.

5 Results

5.1 WMT25 AutoRank

Table 2 reports the official preliminary automatic results for Japanese→Chinese. AutoRank is computed by ensembling multiple metrics (KIWI-XL, GEMBA-ESA-CMDA, GEMBA-ESA-GPT-4.1, MetricX-24 Hybrid-XL, and XCOMET-XL). Our system ranks 4th out of 41 valid submissions, outperforming several large closed models (e.g., GPT-4.1, Claude-4, DeepSeek-V3, Mistral-Medium). Because our selection stage employs metrics related to those in AutoRank, some metric coupling is possible (Pombal et al., 2025); we therefore treat absolute deltas with caution and emphasize the human evaluation below.

5.2 WMT25 Human Evaluation

Table 1 presents the final human evaluation results, adopted from the official WMT25 findings (Kocmi et al., 2025a). As shown in the table, our submission system, Kaze-MT ranked 8th, slightly lagging behind the massive closed LLMs like DeepSeek-V3, CommandA-WMT while still outperforming systems like Mistral-Medium. Notably, the human evaluation presents lower ranking compared to the Autorank, presenting the automatic translation metric still presents unaligned preference as humans.

Japanese→Chinese

Rank	System	Human	AutoRank
1-1	**Human**	-3.5	
2-2	Gemini-2.5-Pro	-4.4	3.3
3-6	Algharb	-5.8	4.3
3-7	Claude-4	-5.9	6.4
3-7	Shy-hunyuan-MT	-6.1	1.0
3-7	GPT-4.1	-6.2	4.5
4-7	Wenyiil	-6.9	4.5
8-10	CommandA-WMT	-7.7	5.2
8-10	DeepSeek-V3	-8.1	6.5
8-13	Kaze-MT	-8.6	3.9
10-13	Mistral-Medium✗	-10.0	6.6
10-13	In2x✗	-10.0	3.0
10-13	Qwen3-235B	-10.9	7.6
14-15	GemTrans	-10.9	6.6
14-15	NTTSU	-11.3	5.9
16-17	Yolu	-12.6	7.1
16-17	TowerPlus-9B[M]	-13.3	11.5
18-18	IRB-MT	-13.9	12.4
19-19	Laniqo	-18.3	11.3

Table 1: The official WMT25 Human Evaluation results adopted from Kocmi et al. (2025a). The human score is the micro-average of human judgements across all domains and double annotations. AutoRank is calculated from automatic metrics as per (Kocmi et al., 2025b). Significance testing is done using a Wilcoxon signed rank test with a p-value threshold of 5%. Ranks from row in two directions until they reach a system that is significantly different. Clusters are created such that they do not overlap with ranks. Systems are either constrained (white), or unconstrained (gray). Systems that do not officially support the language pair are marked with ✗.

6 Discussion

6.1 On Metric Bias and Coupling Effects

When the selection ensemble and the official evaluation share metric families, *metric coupling* can inflate automatic rankings. In our case, AutoRank includes metrics related to our selectors (e.g., KIWI-22 and MetricX variants), which may partially explain why our AutoRank position exceeds our human-evaluation cluster. This is a form of

Japanese-Simplified Chinese									
System Name	LP Supported	Params. (B)	Humeval?	AutoRank ↓	Kiwi-XL ↑	GEMBA-ESA-CMDA ↑	GEMBA-ESA-GPT4.1 ↑	MetricX-24-Hybrid-XL ↑	XCOMET-XL ↑
Shy-hunyuan-MT	✓	7	✓	1.0	0.577	85.1	85.5	-4.2	0.629
In2x		72	✓	3.0	0.624	77.0	77.7	-4.7	0.618
Gemini-2.5-Pro	✓		✓	3.2	0.549	84.8	84.8	-4.6	0.596
Kaze-MT	✓	72	✓	3.8	0.569	81.5	81.8	-4.8	0.605
Algharb	✓	14	✓	4.2	0.547	83.5	84.1	-4.8	0.583
GPT-4.1	✓		✓	4.4	0.549	83.8	84.7	-5.1	0.582
Wenyiil	✓	14	✓	4.5	0.555	81.4	81.9	-4.8	0.591
CommandA-WMT	✓	111	✓	5.1	0.558	80.2	79.7	-4.7	0.575
NTTSU	✓	14	✓	5.8	0.563	77.5	74.8	-4.6	0.577
bb88				6.1	0.551	80.1	78.9	-5.2	0.573
Claude-4	✓		✓	6.2	0.545	82.9	83.7	-5.6	0.556
DeepSeek-V3	✓	671	✓	6.3	0.534	82.9	80.9	-5.1	0.552
Mistral-Medium			✓	6.4	0.546	81.1	81.1	-5.4	0.558
GemTrans	✓	27	✓	6.5	0.556	76.0	74.9	-4.8	0.579
Yolu	✓	14	✓	6.9	0.578	74.6	73.6	-5.0	0.565
Qwen3-235B	✓	235	✓	7.5	0.549	78.4	77.0	-5.4	0.555
CommandA	✓	111		7.6	0.54	79.4	77.6	-5.5	0.556
UvA-MT	✓	12		8.3	0.564	73.9	75.2	-5.6	0.561
TowerPlus-72B[M]	✓	72		9.7	0.537	76.5	75.0	-5.9	0.536
AyaExpanse-32B	✓	32		10.7	0.537	73.2	72.0	-5.8	0.521
Laniqo	✓	9	✓	11.1	0.579	63.1	62.1	-5.4	0.557
TowerPlus-9B[M]	✓	9	✓	11.2	0.535	71.9	69.8	-5.8	0.523
IRB-MT	✓	12	✓	12.1	0.521	72.2	70.4	-6.0	0.509
Gemma-3-27B	✓	27		12.8	0.526	70.4	70.2	-6.2	0.503
Llama-4-Maverick	✓	400		13.1	0.524	71.5	66.1	-6.3	0.518
Qwen2.5-7B	✓	7		13.6	0.524	68.9	67.4	-6.3	0.502
IR-MultiagentMT				13.7	0.523	67.8	68.5	-6.2	0.492
SRPOL		12		13.8	0.56	63.8	62.5	-6.4	0.522
EuroLLM-22B-pre.[M]	✓	22		14.7	0.521	66.4	66.2	-6.3	0.486
AyaExpanse-8B	✓	8		15.5	0.518	65.6	64.4	-6.4	0.472
ONLINE-B	✓			16.2	0.499	63.7	63.2	-6.2	0.472
Gemma-3-12B	✓	12		17.1	0.509	65.0	64.1	-7.1	0.465
CommandR7B	✓	7		18.4	0.496	59.8	58.5	-6.9	0.486
TranssionTranslate				18.8	0.488	59.9	60.6	-6.7	0.45
Llama-3.1-8B		8		20.2	0.507	58.8	57.3	-7.2	0.423
EuroLLM-9B[M]	✓	9		20.8	0.479	59.4	57.2	-7.6	0.461
ONLINE-W				25.2	0.456	52.3	52.9	-7.9	0.387
Mistral-7B		7		32.8	0.445	42.9	43.4	-9.8	0.317
SalamandraTA	✓	8		33.1	0.426	36.5	38.0	-8.6	0.328
ONLINE-G	✓			40.8	0.352	39.5	39.8	-12.1	0.28
NLLB	✓	1		41.0	0.371	35.5	35.8	-12.1	0.303

Table 2: The official WMT25 AutoRank results adopted from Kocmi et al. (2025b). Our submission, Kaze-MT ranked 4th out of 41 valid submissions. Note that our submission is based on the best-of-N reranking using KIWI22, MetricX24-QE-XXL, and Remedy24-QE, thus such approach could deliver biased results when using the same model for reranking and evaluation.

evaluation-on-the-features bias: the system is optimized for the very signals (or close proxies) used to score it.

6.2 Potential discrepancy between human and automatic metrics

Another potential reason is the gap between human judgments and current automatic metrics. Most widely used metrics are black-box models: they output a single overall score without exposing intermediate decisions or confidence. Without explanations, these scores can reflect surface cues (e.g., lexical overlap, length) rather than the properties humans care about (translation accuracy, register). They may also miss context-sensitive errors (tone, pragmatics) and discourse links across sentences.

As a result, a system optimized to rank well under such metrics can improve automatic scores without a matching gain in human preference. This points to the need for more interpretable evaluation models.

7 Conclusion

We presented Kaze-MT, a simple yet competitive TTS system for Japanese–Chinese MT. By pairing diverse zero-shot sampling from a strong LLM with robust QE-based selection, we achieve strong results without any fine-tuning or in-domain resources. The modest gap between AutoRank and human ranking of our submission indicates that evaluation-on-the-features bias may exist, and TTS approach largely depends on the quality and robustness of quality estimation metrics. Reduce metric coupling and improving alignment of quality estimation methods with human preferences remains an important future work.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

References

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of wmt25 general machine translation systems.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264.

Tomoya Mizumoto and Yuji Matsumoto. 2016. Discriminative reranking for grammatical error correction with statistical machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1133–1138.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: Naist at wat2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 35–41.

José Pombal, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation. *arXiv preprint arXiv:2503.08327*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. Investigating test-time scaling with reranking for machine translation.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling. *arXiv preprint arXiv:2504.13630*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*.

NTTSU at WMT2025 General Translation Task

Zhang Yin ♠, **Hiroyuki Deguchi** ♢, **Haruto Azami** ♠, **Guanyu Ouyang** ♠,
Kosei Buma ♠, **Yingyi Fu** ♠, **Katsuki Chousa** ♢, **Takehito Utsuro** ♠
♠University of Tsukuba ♢NTT, Inc.

Abstract

This paper presents the submission of NTTSU for the constrained track of the English–Japanese and Japanese–Chinese at the WMT2025 general translation task. For each translation direction, we build translation models from a large language model by combining continual pretraining, supervised fine-tuning, and preference optimization based on the translation quality and adequacy. We finally generate translations via context-aware MBR decoding to maximize translation quality and document-level consistency.

1 Introduction

We describe our NTTSU translation system in the WMT'25 English–Japanese (En–Ja) and Japanese–Chinese (Ja–Zh) general translation task under the constrained track.

Our translation models are trained on a pretraining large language model (LLM), Qwen3-14B (Qwen Team, 2025). We combine training methods for each translation direction from three training stages: continual pretraining (CPT) (Ke et al., 2023), supervised fine-tuning (SFT) (Zhang et al., 2024), and preference optimization (PO) (Rafailov et al., 2023). In PO, to maximize the translation quality and adequacy, we use two different reward metrics, MetricX-24 (Juraska et al., 2024) and coverage of word alignment between source and target texts (Wu et al., 2024). After training the models, we generate translations using context-aware minimum Bayes risk (MBR) decoding, which maximizes the expected translation quality (Kumar and Byrne, 2004; Eikema and Aziz, 2020) and also utilizes context information of both source and generated target texts, though we use a sentence-level metric (Kudo et al., 2024; Pombal et al., 2024). The following sections show the details of our system.

2 Approaches

2.1 Training

Continual pretraining Continual pretraining (CPT) continues to train LLM models based on the next token prediction as well as pretraining using monolingual corpora (Ke et al., 2023). Let $\mathbf{y} \coloneqq (y_1, y_2, \ldots, y_{|\mathbf{y}|}) \in \mathcal{V}^*$ be a sequence of tokens in a corpus, where $\mathcal{V}^*$ is the Kleene closure of vocabulary $\mathcal{V}$. CPT optimizes the model parameter θ by minimizing the loss function $\mathcal{L}_{\text{CPT}}$ over a monolingual corpus $\mathcal{D}_{\text{CPT}} \coloneqq \{\mathbf{y}_i\}_{i=1}^{|\mathcal{D}_{\text{CPT}}|} \subset \mathcal{V}^*$:

$$\underset{\theta}{\operatorname{argmin}} \sum_{\mathbf{y} \in \mathcal{D}_{\text{CPT}}} \mathcal{L}_{\text{CPT}}(\mathbf{y}; \theta), \tag{1}$$

$$\mathcal{L}_{\text{CPT}}(\mathbf{y}; \theta) \coloneqq -\sum_{t=1}^{|\mathbf{y}|} \log p_\theta(y_t | \mathbf{y}_{<t}). \tag{2}$$

For efficiency, $\mathbf{y}_{[t-c,t)} \coloneqq (y_{t-c}, y_{t-c+1}, \ldots y_{t-1})$ is used instead of $\mathbf{y}_{<t}$ in practice, where $c \in \mathbb{N}$ is a length of a context window. This objective is the same as the pretraining loss of causal language models, i.e., the model is trained to predict the next token y_t under the condition of c context tokens.

Supervised fine-tuning Supervised fine-tuning (SFT) adapts a pretrained model to downstream tasks using labeled data (Zhang et al., 2024). Specifically, given a pretrained model parameter θ, SFT updates it on a labeled dataset $\mathcal{D}_{\text{SFT}} \coloneqq \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{D}_{\text{SFT}}|} \subset \mathcal{V}^* \times \mathcal{V}^*$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the input and its corresponding ground-truth output sequence, respectively, as follows:

$$\underset{\theta}{\operatorname{argmin}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}_{\text{SFT}}} \mathcal{L}_{\text{SFT}}(\mathbf{x}, \mathbf{y}; \theta), \tag{3}$$

$$\mathcal{L}_{\text{SFT}}(\mathbf{x}, \mathbf{y}; \theta) \coloneqq -\log p_\theta(\mathbf{y} \mid \mathbf{x}). \tag{4}$$

This encourages the model to generate outputs that are consistent with the human-annotated targets.

Proceedings of the Tenth Conference on Machine Translation, pages 657–665
November 8-9, 2025

Preference optimization Preference optimization (PO) aims to align a trained model with preferences. One of the major PO algorithms is direct PO (DPO), which uses pairwise comparison data instead of explicit reward models (Rafailov et al., 2023). Let $\mathcal{D}_{\text{PO}} \coloneqq \{(\mathbf{x}_i, \mathbf{y}_i^+, \mathbf{y}_i^-)\}_{i=1}^{|\mathcal{D}_{\text{PO}}|} \subset \mathcal{V}^* \times \mathcal{V}^* \times \mathcal{V}^*$ be a triplet dataset that consists of a prompt $\mathbf{x}$ and its corresponding output pairs $(\mathbf{y}^+, \mathbf{y}^-)$, where $\mathbf{y}^+$ is preferred over $\mathbf{y}^-$ according to human feedback or a reward function, i.e., $\mathbf{y}^+ \succeq \mathbf{y}^-$. PO tunes a model θ by minimizing a pairwise loss that encourages the model to generate $\mathbf{y}^+$ rather than $\mathbf{y}^-$. We minimize the following objective function that incorporates adaptive rejection (Xu et al., 2025) into SimPO, a variant of DPO (Meng et al., 2024):

$$\operatorname*{argmin}_{\theta} \sum_{(\mathbf{x},\mathbf{y}^+,\mathbf{y}^-)\in\mathcal{D}_{\text{PO}}} \mathcal{L}_{\text{PO}}(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-; \theta), \quad (5)$$

$$\begin{aligned}\mathcal{L}_{\text{PO}}(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-; \theta) \coloneqq \\ -\log \sigma \left(r(\mathbf{x}, \mathbf{y}^+) - \tau_\theta(\mathbf{y}^+, \mathbf{y}^-) r_\theta(\mathbf{x}, \mathbf{y}^-) - \gamma\right) \\ + \alpha \log p_\theta(\mathbf{y}^+|\mathbf{x}), \quad (6)\end{aligned}$$

where $\alpha \in \mathbb{R}$ is a weight of the behavior cloning regularizer, $\gamma \in \mathbb{R}$ is a reward margin between $\mathbf{y}^+$ and $\mathbf{y}^-$. Note that $r_\theta(\mathbf{x}, \mathbf{y})$, $\tau_\theta(\mathbf{y}^+, \mathbf{y}^-)$, and $z_\theta(\mathbf{y}^+, \mathbf{y}^-)$ are defined as follows:

$$r_\theta(\mathbf{x}, \mathbf{y}) \coloneqq \frac{\beta}{|\mathbf{y}|} \log p_\theta(\mathbf{y}|\mathbf{x}), \quad (7)$$

$$\tau_\theta(\mathbf{y}^+, \mathbf{y}^-) \coloneqq \min\left(e^{\eta \cdot z_\theta(\mathbf{y}^+, \mathbf{y}^-)} - 1, 1\right), \quad (8)$$

$$z_\theta(\mathbf{y}^+, \mathbf{y}^-) \coloneqq \left|\frac{\log p_\theta(\mathbf{y}^+|\mathbf{x})}{|\mathbf{y}^+|} - \frac{\log p_\theta(\mathbf{y}^-|\mathbf{x})}{|\mathbf{y}^-|}\right|, \quad (9)$$

where $\beta \in \mathbb{R}$ and $\eta \in \mathbb{R}$ are hyperparameters.

Stepwise preference optimization Stepwise PO (Wachi et al., 2024) is an extension of PO designed to align models with multiple preference metrics. It optimizes the model using multiple preferences sequentially, where each stage focuses on a distinct preference objective. Consequently, by chaining multiple preference optimization stages, the model incrementally aligns with multiple-perspective preferences.

2.2 Decoding

We generate translations via context-aware minimum Bayes risk (MBR) decoding, which leverages sentence-level metrics for MBR decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004; Eikema and Aziz, 2020) yet utilizes both source and generated target context information.

Algorithm 1: Context-aware MBR decoding

Given : Translation model θ, utility function u, the number of hypotheses $|\mathcal{H}|$, and the context size $C \in \mathbb{N}$.
Input : Source document $\mathbf{X} \coloneqq (\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathbf{X}|})$ where $\mathbf{x}_i \in \mathcal{V}^*$ is the i-th source sentence.
Output : Target document $\mathbf{Y} \coloneqq (\mathbf{y}_1, \ldots, \mathbf{y}_{|\mathbf{Y}|})$.

$\mathbf{Y} \leftarrow \phi$
Create queues: $\mathbf{C}_x \leftarrow \phi$ and $\mathbf{C}_y \leftarrow \phi$
for $i \leftarrow 1 \ldots |\mathbf{X}|$ **do**
  Enqueue$(\mathbf{C}_x, \mathbf{x}_i)$
  // $\mathcal{H}$ is a multiset of hypotheses.
  $\mathcal{H} \leftarrow \{\mathbf{h}_k \sim p(\mathbf{y}_i|\mathbf{C}_x, \mathbf{C}_y; \theta)\}_{k=1}^{|\mathcal{H}|}$
  // We use the same candidate set for hypotheses and pseudo-references.
  $\hat{\mathbf{y}}_i \leftarrow \operatorname{argmax}_{\mathbf{h}\in\mathcal{H}} \frac{1}{|\mathcal{H}|} \sum_{k=1}^{|\mathcal{H}|} u(\mathbf{h}, \mathbf{h}_k)$
  // $\circ$ denotes concatenetion.
  $\mathbf{Y} \leftarrow \mathbf{Y} \circ \hat{\mathbf{y}}_i$
  Enqueue$(\mathbf{C}_y, \hat{\mathbf{y}}_i)$
  while $|\mathbf{C}_x| > C$ **do**
    Dequeue$(\mathbf{C}_x)$
  while $|\mathbf{C}_y| > C$ **do**
    Dequeue$(\mathbf{C}_y)$
return $\mathbf{Y}$

MBR decoding The goal of MBR decoding is to find a translation that maximizes the expected utility rather than the output probability (Goel and Byrne, 2000; Kumar and Byrne, 2004). The objective is formally defined as follows:

$$\mathbf{y}^\star_{\text{MBR}} \coloneqq \operatorname*{argmax}_{\mathbf{h}\in\mathcal{V}^*} \mathop{\mathbb{E}}_{\mathbf{y}\sim\Pr(\cdot|\mathbf{x})}[u(\mathbf{h}, \mathbf{y})], \quad (10)$$

where $\Pr(\cdot|\mathbf{x})$ is the true probability of human translation and $u\colon \mathcal{V}^* \times \mathcal{V}^* \to \mathbb{R}$ is a utility function that evaluates a hypothesis under the given reference $\mathbf{y}$ and satisfies $\mathbf{h}^+ \succeq \mathbf{h}^- \iff u(\mathbf{h}^+, \mathbf{y}) \geq u(\mathbf{h}^-, \mathbf{y})$. Since searching over $\mathcal{V}^*$ and calculating the expectation over the output space are infeasible, the objective of MBR decoding is approximated by the Monte Carlo (MC) estimation (Eikema and Aziz, 2020, 2022). We denote a hypothesis set by $\mathcal{H} \subset \mathcal{V}^*$. The MBR decoding with the MC estimation is calculated as follows:

$$\mathbf{y}_{\text{MBR}} \coloneqq \operatorname*{argmax}_{\mathbf{h}\in\mathcal{H}} \frac{1}{|\hat{\mathcal{Y}}|} \sum_{\mathbf{y}\in\hat{\mathcal{Y}}} u(\mathbf{h}, \mathbf{y}), \quad (11)$$

where $\hat{\mathcal{Y}} \coloneqq \{\mathbf{y}_i \sim p_\theta(\mathbf{y}|\mathbf{x})\}_{i=1}^{|\hat{\mathcal{Y}}|}$ is a multiset, a.k.a. bag, of pseudo-references, translation samples drawn from the output probability of translation model θ. Typically, hypotheses are also used as pseudo-references.

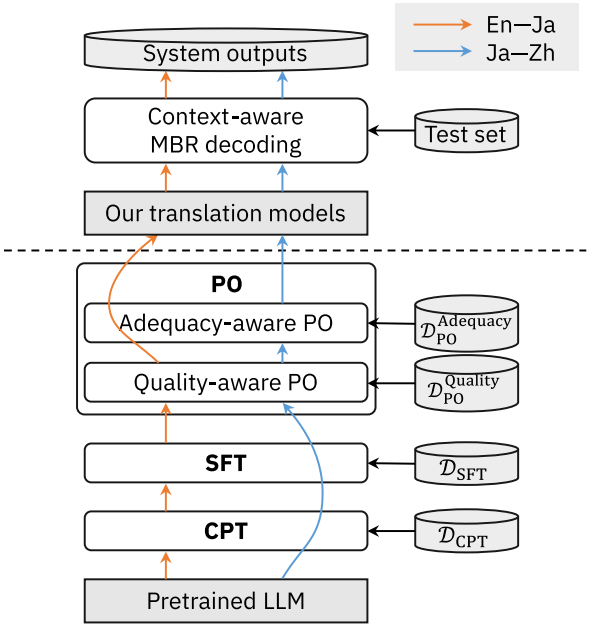


Figure 1: Overview of our translation system.

Method	En–Ja	Ja–Zh
Continual pretraining (CPT)	✓	✗
Supervised fine-tuning (SFT)	✓	✗
Preference optimization (PO)	—	—
Quality-aware PO	✓	✓
Adequacy-aware PO	✗	✓
Context-aware MBR decoding	✓	✓

Table 1: List of methods we employed.

Hyperparameter	CPT	SFT
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) (Loshchilov and Hutter, 2019)	
Learning rate	2.5×10^{-5}	1×10^{-6}
Scheduler	cosine	inverse square root
Warmup ratio	1%	1%
Weight decay	0.1	0.1
Gradient clip	1.0	1.0
Epoch	1	3
Batch size	1,024 chunks	64 sentence pairs
Chunk size	2,048 tokens	N/A
Accelarator	DeepSpeed ZeRO-2 (Rasley et al., 2020)	
Precision	bfloat16	bfloat16

Table 2: Hyperparameters of CPT and SFT.

Context-aware MBR decoding For document-level translation, we extend MBR decoding to a context-aware method. However, most automatic evaluation metrics, which are used for the utility function, are designed for sentence-level metrics. To bridge this gap, we determine output translations for each sentence by MBR decoding with a sentence-level utility and add the generated translation to the context, similar to Kudo et al. (2024) and Pombal et al. (2024). Algorithm 1 shows our decoding algorithm, which autoregressively generates sentence-level translations at each step. Since the WMT'25 general translation task provides the source document without sentence segmentation, we first apply a sentence segmenter before running our decoding algorithm. In Line 5, hypotheses are sampled given the source and target context sentences, i.e., $\mathbf{C}_x$ and $\mathbf{C}_y$, respectively. The source context includes the current source sentence $\mathbf{x}_i$ as well as the preceding ones, while the target context consists only of previously generated target sentences. Accordingly, the model focuses on the current sentence $\mathbf{x}_i$ and generates its corresponding target sentence $\mathbf{y}_i$ under the given contexts, naturally. The hyperparameter $C \in \mathbb{N}$ denotes the size of the context queues. Rather than using a fixed number of sentences, we set it based on the paragraph size, i.e., a variable number of sentences depending on the dataset format.

3 Submission System

We train En–Ja and Ja–Zh translation models from a pretrained LLM, Qwen3-14B (Qwen Team, 2025). According to our preliminary experiments and subjective judgment, we selected the combinations of training methods. Finally, we generate translations via context-aware MBR decoding. We show the system overview in Figure 1 and Table 1.

3.1 Continual pretraining

We perform the bilingual CPT only for the En–Ja model. For the training data of CPT, we use JParaCrawl v3.0 (Morishita et al., 2022) and filter it into 20.8M sentence pairs using LEALLA-large (Mao and Nakagawa, 2023). We create training examples following Kondo et al. (2024). The hyperparameters of CPT are listed in Table 2.

3.2 Supervised fine-tuning

Similar to CPT, we conduct supervised fine-tuning (SFT) only for the En–Ja model. For the training data of SFT, we use the development and test sets of the WMT'20 translation task (Barrault et al., 2020) and FLoRes-200 (NLLB Team et al., 2022), along with the train set of the Kyoto Free Translation Task (KFTT) (Neubig, 2011). For the development set, we use the test set of the WMT'21 translation task (Akhbardeh et al., 2021). The hyperparameters of SFT are also listed in Table 2.

3.3 Preference optimization

To maximize translation quality and adequacy, we perform PO with two reward metrics. For En–Ja, we apply the quality-aware PO on top of the model trained via SFT. For Ja–Zh, we apply both quality- and adequacy-aware PO through stepwise PO. Note that we do not apply CPT and SFT to the Ja–Zh model as listed in Table 1; thus, we directly tune the pretrained Qwen3-14B.

Quality-aware PO To improve translation quality, we employ an automatic evaluation metric that highly correlates with human assessments for creating the preference data. Specifically, we first randomly sample 20,000 source sentences from the NewsCrawl corpus (Kocmi et al., 2024), and generate translations for each source sentence using two LLMs, Qwen3-32B and Aya-Expanse-32B (Dang et al., 2024). To obtain high-quality translations efficiently, we employ COMET[1] (Rei et al., 2022a) -based MBR decoding using 64 hypotheses sampled via epsilon sampling with $\varepsilon = 0.02$ (Freitag et al., 2023). These high-quality translations and the baseline translations, generated via beam search from Qwen3-14B, are then compared using MetricX-24-XXL (Juraska et al., 2024). Among the outputs from the three models, we label the highest-quality translation as the preferred, i.e., chosen, instance and the lowest-quality translation as the non-preferred, i.e., rejected, instance. From these paired instances with each source sentence, we construct the training dataset for quality-aware PO. Finally, we train the model by optimizing Equation (5) on the created preference data with $\alpha = 1.0, \beta = 0.2, \eta = 1.5, \gamma = 0.005$.

Adequacy-aware PO To mitigate hallucination and omission, i.e., overgeneration and undergeneration, we also employ the word alignment-based preference metric (Wu et al., 2024) for Ja–Zh. For the preference data, we randomly sample 10,000 source sentences from CCAligned (El-Kishky et al., 2020), where each sentence has at least 15 characters. To label the preference data, we use the coverage score obtained via word alignment calculated by WSPAlign[2] (Wu et al., 2023). Apart from these two modifications, we follow the same procedure as in the quality-aware PO, but with different hyperparameters: $\alpha = 1.0, \beta = 0.01, \eta = 1.5, \gamma = 0.005$.

[1] https://huggingface.co/Unbabel/wmt22-comet-da

[2] https://huggingface.co/qiyuw/WSPAlign-mbert-base

3.4 Prompt templates

We basically use the following template that turns on the `continue_final_message` (CFM) option defined in the tokenizers of Huggingface transformers (Wolf et al., 2020):

```
<|im_start|>user
Translate this from English to Japanese:
English: ...<|im_end|>
<|im_start|>assistant
<think>
</think>
Japanese:
```

We call this "CFM" template. The CFM template inserts the target language name with a colon into the last of the assistant chat and does not close it. Hence, the model naturally generates a target text following the target language name.

However, in our preliminary Ja–Zh translation experiments, we observed that generated texts with the CFM template are often collapsed due to hallucinations. Thus, we change the inference template to the below "AGP" template, which enables the `add_generation_prompt` (AGP) option instead of the `continue_final_message` option:

```
<|im_start|>user
Translate this from English to Japanese:
English: ...
Japanese:<|im_end|>
<|im_start|>assistant
<think>
</think>
```

Although there is a slight difference between training and inference, we employ this method because hallucinations decrease in Ja–Zh.

To summarize, we train both En–Ja and Ja–Zh models with the CFM template, and generate translations with CFM for En–Ja and AGP for Ja–Zh.

3.5 Decoding

In decoding, we use MetricX-24-XXL (Juraska et al., 2024) for the utility function u. During decoding, we generate 64 translation candidates via epsilon sampling with $\varepsilon = 0.02$ (Freitag et al., 2023) and use them for both hypotheses and pseudo-references. We split the source documents into sentences using segment-any-text[3] (Frohmann et al., 2024). We use at most one previous paragraph as context, i.e., the target context includes a preceding generated paragraph and generated sentences until the current focused sentence.

[3] https://huggingface.co/segment-any-text/sat-12l-sm

CPT	SFT	PO: Quality	PO: Adequacy	MTX24↓	XCMT↑	KIWI22↑
✗	✗	✗	✗	4.44	79.89	83.04
			✓	4.39	80.28	82.97
		✓	✗	4.32	80.60	83.08
			✓	4.44	79.67	83.08
	✓	✗	✗	4.82	76.58	81.89
			✓	4.67	77.82	82.52
		✓	✗	4.32	80.82	83.05
			✓	4.47	79.31	83.06
✓	✗	✗	✗	Failed	Failed	Failed
			✓	5.35	72.97	81.04
		✓	✗	4.30	81.02	**83.15**
			✓	4.50	79.81	83.09
	✓	✗	✗	4.97	75.82	81.36
			✓	4.79	77.10	82.07
		✓	✗	**4.29**	**81.27**	83.11
			✓	4.54	79.66	82.99

Table 3: Comparisons of training methods on the WMT'24 En–Ja test set. The **bold** font indicates the best scores in each metric. The green highlighted rows indicate the setting of our submission system.

CPT	SFT	PO: Quality	PO: Adequacy	MTX24↓	XCMT↑	KIWI22↑
✗	✗	✗	✗	3.51	73.55	73.26
			✓	3.44	73.75	73.12
		✓	✗	3.46	74.03	73.12
			✓	**3.43**	**74.36**	**73.38**
✗	✓	✗	✗	4.17	70.20	72.38
			✓	4.08	70.96	72.32
		✓	✗	3.53	73.17	73.10
			✓	3.54	73.26	73.15
✓	✗	✗	✗	Failed	Failed	Failed
			✓	Failed	Failed	Failed
		✓	✗	3.92	66.81	71.73
			✓	4.06	66.71	72.00
✓	✓	✗	✗	5.43	63.97	70.63
			✓	4.38	68.10	71.42
		✓	✗	3.57	71.28	73.07
			✓	3.65	71.14	73.10

Table 4: Comparisons of training methods on the WMT'24 Ja–Zh test set. The **bold** font indicates the best scores in each metric. The green highlighted rows indicate the setting of our submission system.

4 Experiments

4.1 Ablation study of training methods

We investigate the effects of each training method.

Setup We compare the combination of training methods: CPT, SFT, quality-aware PO, and adequacy-aware PO. We train models with the same training data and hyperparameters as our submission system, as described in Section 4.1, except for the differences noted below. In Ja–Zh, we use the same hyperparameters as listed in Table 2 for both CPT and SFT. For the training data of CPT in Ja–Zh, we use the parallel corpora listed in "WMT 2025 Translation Task Training Data"[4]. We filter them to retain only those with CometKiwi-22 (Rei et al., 2022b) scores between 0.5 and 0.88, and then clean them using bifixer (Ramírez-Sánchez et al., 2020). In both En–Ja and Ja–Zh, the source sides of training examples are shared between SFT and adequacy-aware PO. The translation quality is evaluated on MetricX-24-XXL (MTX24) (Juraska et al., 2024), xCOMET-XXL (XCMT) (Guerreiro et al., 2024), and CometKiwi-22 (KIWI22) (Rei et al., 2022b) in the test sets of WMT'24 En–Ja and Ja–Zh translation tasks (Kocmi et al., 2024).

[4] https://www2.statmt.org/wmt25/mtdata/

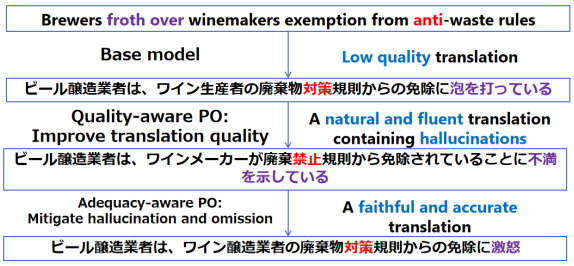


Figure 2: Examples of stepwise preference optimization (PO). Purple and red texts highlight corresponding phrases in the source text and its translations. Blue text provides descriptive labels for each step.

Results The results of automatic evaluation on the WMT'24 test sets are demonstrated in Table 3 and Table 4. In the tables, "Failed" indicates that it failed to generate translations due to hallucinations or critical errors, and it cannot be evaluated. As demonstrated in Table 3 and Table 4, the configuration of our submission system achieved the best MetricX-24-XXL and xCOMET-XXL scores in both En–Ja and Ja–Zh. In addition, we also confirmed through subjective judgment that these models have successfully generated the highest-quality translations compared to other settings. Accordingly, we selected these combinations of training methods for each translation direction.

Figure 2 shows translation examples of the same sentence from the WMT24 test set, generated by

the base model (Qwen3-14B), the model after an initial training step with Quality-aware PO, and the model after a subsequent step with Adequacy-aware PO. As shown in the figure, the initial translation from the base model is relatively low quality, incorrectly translating "froth over." After training with Quality-aware PO, the translation becomes more fluent and natural overall. However, hallucinations occur during translation—for example, rendering "froth over," which expresses anger, as merely expressing dissatisfaction, and interpreting "anti," which denotes countermeasures, as "prohibit." In contrast, the model trained with Adequacy-aware PO produces a translation that is accurate and faithful to the source text.

It is noteworthy that among the three translations, the one from the Quality-aware PO model achieved the best MetricX-24-XXL score. From up to below, the MetricX-24-XXL scores are 6.34, 4.99, and 5.54. This indicates that even advanced metrics such as MetricX-24-XXL may assign better scores to fluent and natural translations that contain hallucinations than to factually accurate but less fluent ones.

4.2 Comparison of decoding strategy

We evaluate our decoding algorithm using the final submission system by comparing it with a baseline context-aware MAP decoding.

Setup We use the WMT'25 En–Ja and Ja–Zh translation task and evaluate the translation quality of decoding methods using reference-free quality estimation (QE) models, MetricX-23-QE-XXL (Juraska et al., 2023), MetricX-24-XXL[5] (Juraska et al., 2024), and CometKiwi-23-XXL (Rei et al., 2023). We compare our decoding algorithm with context-aware MAP decoding, which employs a beam search with a beam size of 5. The context sizes of both methods are at most one previous paragraph, as described in Section 3.5.

For evaluation, we first apply a sentence segmenter[6] (Frohmann et al., 2024) to each source and target paragraph and compute the scores across all pairs of source and target sentences for each paragraph. Then, we compute the score alignment that maximizes the total scores. Finally, the document-level QE scores are calculated by averaging the paragraph-level QE scores.

[5]MetricX-24 is a hybrid reference-based/-free metric, so we use it as a reference-free QE model in this evaluation.

[6]https://huggingface.co/segment-any-text/sat-12l-sm

Direction	Decoding	MTX23↓	MTX24↓	KIWI23↑
En–Ja	MAP	3.4	4.9	74.7
	MBR	**3.0**	**4.2**	**77.2**
Ja–Zh	MAP	4.0	4.9	63.1
	MBR	**3.5**	**4.7**	**64.8**

Table 5: Reference-free quality estimation scores on the WMT'25 test set. The **bold** font indicates the best scores in each translation direction. The green highlighted rows indicate the setting of our submission system.

Results Table 5 demonstrates the translation quality of decoding methods. The table shows that MBR decoding consistently outperformed MAP decoding across all metrics, even though we used only MetricX-24-XXL for the utility function. One reason for these results is that MAP decoding tends to propagate translation errors, including hallucinations, whereas MBR decoding carefully selects translations based on the expected utility computed from the evaluation metric and other translation samples, thereby mitigating the generation of pathological sequences.

5 Conclusion

We built our system on the WMT'25 general translation task in En–Ja and Ja–Zh. Our models were trained with the combinations of CPT, SFT, and stepwise PO based on the quality- and adequacy-aware rewards, for each translation direction. To maximize the translation quality and document-level consistency, we generated translations via context-aware MBR decoding.

In document-level translation, we observed that LLMs are more likely to generate collapsed hallucination texts. To mitigate this issue, we employed adequacy-aware PO. Nevertheless, in some cases, the models still failed to generate translations. We hope to further improve hallucination mitigation in document translation.

Limitations

Metric bias We used MetricX-24-XXL for the preference data creation in PO and the utility function of MBR decoding, which heavily relied on a single metric. Thus, our system may be affected by the metric bias.

Domain adaptation We built a single system for each translation direction, regardless of domains,

while the WMT'25 general translation task contains multiple domains. By considering domain-specific knowledge and preferences, further improvements in translation quality can be expected.

Multimodal translation In the speech domain, original videos are also provided in addition to plain texts transcribed by an automatic speech recognition (ASR), but we did not use them. This means that ours is a cascade-style speech-to-text or video-to-text translation in the speech domain. By utilizing the original videos and audio, we can expect to suppress the propagation of errors caused by an ASR system.

Acknowledgements

This work was done mainly under a collaborative research agreement between NTT and Tsukuba University. Additionally, this work partially used computational resources of Pegasus provided by the Multidisciplinary Cooperative Research Program in the Center for Computational Sciences, University of Tsukuba.

Author Contributions

Zhang Yin applied PO, conducted translation experiments as described in Section 4.1 and other preliminary experiments, and selected the submission system.
Hiroyuki Deguchi conducted context-aware MBR decoding, as described in Section 2.2, and experiments regarding decoding strategies as shown in Section 4.2.
Haruto Azami applied CPT in En–Ja and generated the preference data in En–Ja for PO.
Guanyu Ouyang applied SFT in En–Ja and selected the submission system.
Kosei Buma collected and cleaned up the Ja–Zh dataset for SFT and PO via translation scoring and deduplication.
Yingyi Fu participated in discussions and reviewed system performance.
Katsuki Chousa provided advice on model development and decoding.
Takehito Utsuro built and managed our team.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), pages 5960–5969, Online. Association for Computational Linguistics.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11908–11941, Miami, Florida, USA. Association for Computational Linguistics.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Minato Kondo, Takehito Utsuro, and Masaaki Nagata. 2024. Enhancing translation accuracy of large language models through continual pre-training on parallel data. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 203–220, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, pages 124198–124235. Curran Associates, Inc.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022. JParaCrawl v3.0: A large-scale English-Japanese parallel corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti

Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jose Pombal, Sweta Agrawal, and André Martins. 2024. Improving context usage for translating bilingual customer support chat with large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 993–1003, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Akifumi Wachi, Thien Q. Tran, Rei Sato, Takumi Tanabe, and Youhei Akimoto. 2024. Stepwise alignment for constrained language model policy optimization. In *Advances in Neural Information Processing Systems*, volume 37, pages 104471–104520. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024. Word alignment as preference for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3223–3239, Miami, Florida, USA. Association for Computational Linguistics.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11084–11099, Toronto, Canada. Association for Computational Linguistics.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey.

A* decoding - Fast, precise and diverse decoding for LLMs Samsung R&D Institute Poland participation in WMT2025

Team Name: SRPOL

Adam Dobrowolski, Paweł Przewłocki, Dawid Siwicki
Samsung R&D Institute, Warsaw, Poland
{a.dobrowols2@samsung.com, p.przewlocki@partner.samsung.com, d.siwicki@samsung.com}

Abstract

This work presents an innovative decoding approach utilizing the A* algorithm, which generates a diverse and precise set of translation hypotheses. Subsequent reranking through the Noisy Channel Model Reranking and Quality Estimation selects the best among these diverse hypotheses, leading to a significant improvement in translation quality. This approach achieves up to a 0.5-point reduction in the MetricX-24 score and a 1.5-point increase in the COMET score.

The A* decoding algorithm is model-agnostic and could be applied to decoding in LLMs as well as classic transformer architectures. The experiment shows that by using freely available open source MT models, it is possible to achieve translation quality comparable to the best online translators and LLMs using a single 32GB GPU card.

1 Introduction

The final stage of inference in language models is decoding, which involves generating text based on calculated token probabilities. The most commonly used approach is autoregressive decoding, which generates tokens sequentially based on the source text and the text produced thus far. There are three primary strategies for autoregressive decoding: greedy, beam, and sampling. Each of them has its flaws that A* resolves.

2 Decoding Strategies

Greedy Decoding selects the most probable token at each step, prioritizing speed and simplicity but often producing suboptimal results due to its limited exploration of alternative paths.

Sampling introduces randomness by selecting tokens based on their probability distribution, but the generated hypotheses can be overly random, resulting in translations that are not always optimal.

Beam Search maintains multiple high-probability sequences (beams) simultaneously, enhancing output quality at a modest computational cost proportional to the beam width. This algorithm has been widely used in machine translation.

Recent advances in large language models (LLMs) have demonstrated their superiority over traditional encoder-decoder transformers (Vaswani et al., 2017). However, because of their primary purpose, LLMs are designed for rapid text generation and not for quality-oriented beam search. Beam search in LLMs takes much more time than in classic transformers. For example, from our experiments, the beam search using vLLM[1] on EuroLLM-9B (Martins et al., 2025) takes about 30 seconds to decode with beam of width 10 for just one sentence. Within the same time, we can use sampling to generate as many as 800 diverse hypotheses, which later assessed by QE give great improvement over greedy decoding or beam search. In the following, we propose a novel A* decoding algorithm for LLMs that combines the speed of beam search in classic transformers with the diversity of sampling.

3 A* Decoding

Let's consider decoding as the process of searching for the optimal path within a tree of all possible sentences generated by a model, where the chosen path represents the best translation of the source sentence. Typically, there are no clear criteria for evaluating the quality of the generated sequence, so we must rely on proxy metrics derived from available data to assess its quality. One of the most effective and straightforward proxy metrics is the average probability of all tokens generated. This metric guides the algorithm to select the most probable token at each step of the output generation

[1] https://docs.vllm.ai/

Proceedings of the Tenth Conference on Machine Translation, pages 666–670
November 8-9, 2025

```
def a_star_decoding(source, max_cands, hope_level):
   queue = ReversedPriorityQueue()
   queue.put((0, []))
   results = []
   for iteration in range(max_cands):
      if queue.empty(): break
      old_f_score, prefix = queue.pop()

      new_trn = LLM.generate(source, prefix)
      results.append(new_trn)
      if iteration == 0:
         def_f_score  = sum(new_trn.tokens) / len(new_trn.tokens)

      for idx, token in enumerate(new_trn.tokens[len(prefix):]):
         idx += len(prefix)
         for alt_token in token.alternative_tokens:
            g_score = sum(new_trn.tokens[:idx]) + alt_token.logprob
            h_score = sum(new_trn.tokens[idx+1:])
            f_score = (g_score + h_score) / len(new_trn.tokens) + hope_level
            if f_score > def_f_score:
               queue.put((f_score, new_trn.tokens[:idx] + [alt_token]))
   return results
```

Listing 1: Python-like pseudocode of the A* decoding

process. The simplest approach, known as greedy decoding, follows this principle. However, the results of greedy decoding may be suboptimal, as it overlooks many potential paths. In contrast, A* decoding allows the algorithm to explore beyond just the most probable next token, selecting tokens that offer the potential for discovering a more probable path by the end of the process. A similar concept has already been explored for Statistical Machine Translation (SMT) with long paragraphs (Och et al., 2001).

3.1 A* algorithm

A* algorithm uses priority queue to explore graph nodes, prioritizing those with the lowest estimated total cost.

For each node n, it computes:

• $g(n)$: exact cost from the start node to n.

• $h(n)$: heuristic estimate of the cost from n to the end of sentence.

• $f(n) = g(n) + h(n)$: estimated total cost of the path through n.

3.2 A* decoding description

We propose the following adaptation of the A* algorithm to decoding in LLMs. The probability of the remaining sequence is the probability of the generated token (as in beam-search) plus probability of the sequence following the new token. Assuming that the remaining sequence could have a slightly higher probability than the base sequence, we can estimate the total probability of a new path. This is done by adding the probability of the new token, adjusted by a small constant to account for the potential increase of total probability.

The initial step of the algorithm is generation of the default greedy translation for the source sentence. We then enhance this baseline by employing the A* algorithm for further exploration. For each token generated, we select alternative tokens at its position and calculate the estimated total cost, f(n), as outlined below:

• $g(n)$ (actual cost) - normalized logarithm probability of a prefix with alternative token.

• $h(n)$ (estimate cost) - normalized logarithm probability of the default suffix plus some constant "hope_level" that assumes that the actual probability of the rest of sentence may be bigger than the default. The "hope_level" should be chosen experimentally. High enough to ensure that $h(n)$ remains admissible (does not overestimate the actual cost), but not too high to avoid considering highly improbable alternative tokens.

The pseudocode for the algorithm is presented in Listing 1.

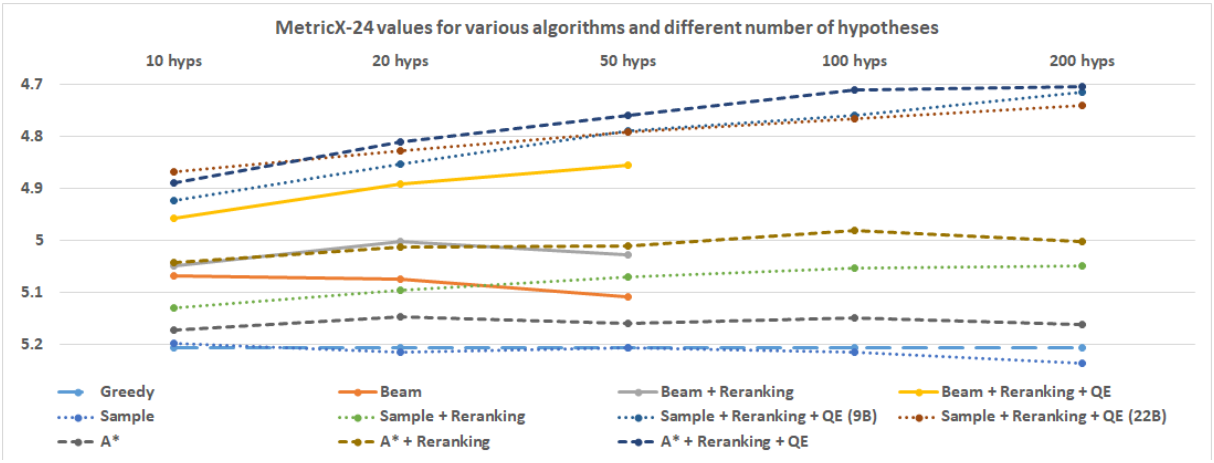


Figure 1: Comparison of A* with other algorithms for the English-to-Czech direction.

Method	10 hyps	20 hyps	50 hyps	100 hyps	200 hyps	time/200hyps
Greedy	5.2064	5.2064	5.2064	5.2064	5.2064	0.031
Beam	5.0677	5.0740	5.1088			3.330
Beam + Reranking	5.0492	5.0021	5.0279			3.330
Beam + Reranking + QE	4.9586	4.8920	4.8555			3.330
Sample	5.1992	5.2149	5.2077	5.2146	5.2363	0.040
Sample + Reranking	5.1303	5.0962	5.0709	5.0539	5.0493	0.042
Sample + Reranking + QE (9B)	4.9233	4.8545	4.7894	4.7604	4.7151	0.047
Sample + Reranking + QE (22B)	**4.8678**	4.8288	4.7919	4.7660	4.7410	0.047
A*	5.1722	5.1468	5.1590	5.1494	5.1612	0.024
A* + Reranking	5.0439	5.0132	5.0111	4.9811	5.0028	0.025
A* + Reranking + QE	4.8889	**4.8106**	**4.7601**	**4.7110**	**4.7049**	0.029

Table 1: Comparison of MetricX-24 scores on the WMT24++ test set for the English-to-Czech translation direction.

4 Reranking

Following the generation of a set of translation hypotheses using the above described algorithm, we select the optimal hypothesis through a modified Noisy Channel Model Reranking approach, as described by (Yee et al., 2019). This reranking method enhances machine translation by reordering candidate translations during the decoding process. It integrates three components: a direct translation model, $P(T|S)$, which predicts the target sentence T given the source sentence S; a channel model, $P(S|T)$, which assesses the likelihood of the source sentence given the target; and a language model, $P(T)$, which evaluates the fluency of the target sentence. The algorithm ranks candidates by computing a weighted combination of these probabilities, ensuring the selection of the most accurate and fluent translation. The algorithm scores candidates using a weighted combination of these probabilities:

• $S(T|S)$=$P(T|S)$+$\lambda_1$log $P(S|T)$+$\lambda_2$log $P(T)$

where λ_1 and λ_2 are tunable weights. Top candidates from the direct model are reranked based on this score, prioritizing translations that balance fidelity to the source and fluency in the target. For our submission in WMT2025 we applied the following scoring:

• $P(T|S)$ - calculated by weighted sum of probabilities of EuroLLM-9B, NLLB-3.3 (Team et al., 2022) combined with Unbabel/comet wmt23-cometkiwi-da-xl score. (Rei et al., 2020)

• $P(S|T)$ - calculated by NLLB-3.3.

• $P(T)$ - not used - for compliance with constrained path.

4.1 Hallucination detection

Before reranking, we remove hypotheses that appear to be outliers or hallucinations. For each source sentence, we calculate the standard deviation of the probability scores across all hypotheses for each scoring model. Hypotheses with scores below the mean minus one standard deviation are filtered out.

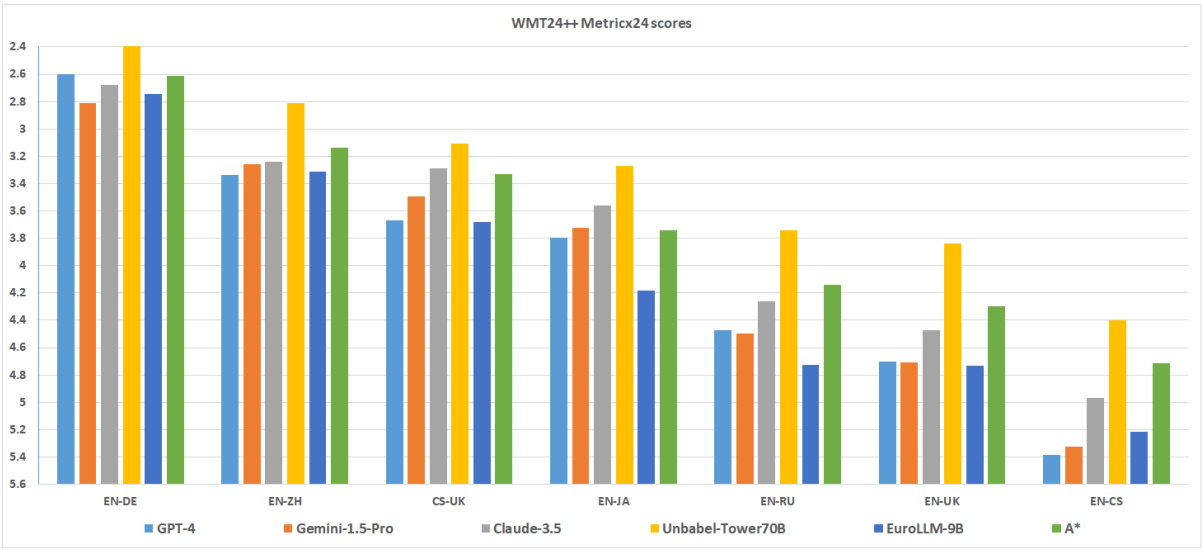


Figure 2: Comparison of presented solution (green) to leading LLMs.

5 Results

5.1 Comparison with different decodings

The chart on Figure 1 presents the quality of various algorithms across different number of generted hypotheses. The Y-axis represents reference-based MetricX-24 scores (Juraska et al., 2024) on wmt24++ testset (Deutsch et al., 2025) for the English-to-Czech translation direction. The X-axis indicates the number of hypotheses generated per source sentence. Table 1 presents exact values for the chart. The values for sampling vary with each run. The values in the table represent the mean of multiple runs. Performance times were measured in seconds on a single A100-80GB GPU using the vLLM framework on EuroLLM-9B. 80GB of GPU memory is not necessary. We have successfully run the above tests on A100-40GB and V100-32GB with slower decoding time, but the same quality.

The baseline performance, established through greedy decoding, achieves a score of 5.2064. Pure beam search, utilizing a beam width of only 10 hypotheses, yields a better score of 5.0667, though it requires approximately 30 seconds to decode a single sentence. Despite its computational complexity, beam search remains the best algorithm for decoding without reranking. The introduction of reranking shifts the advantage to the A* algorithm, which delivers an improved score of 5.0062, representing a 0.2-point improvement over the baseline. Incorporating Quality Estimation (QE) further enhances performance, boosting results by approximately 0.2 to 0.3 points above the reranked results. In this scenario, A* remains the top-performing algorithm, achieving the best score of 4.7049. When all techniques are combined, the overall improvement ranges from approximately 0.2 to 0.5 points on the reference-based MetricX-24 compared to the baseline.

It may seem surprising that A* decoding requires less time than sampling for a single source sentence, but this number comes from total time divided by maximum number of hypotheses - 200. Sampling always generates a fixed number of hypotheses, often including duplicates. In contrast, A* decoding halts once it can no longer identify distinct hypotheses, especially for short segments, resulting in a shorter total processing time compared to sampling. E.g. for beam of 200 A* decoding on WMT24++ generates only about 100 hypotheses on average.

5.2 Comparison with other LLMs

Figure 2 presents automatic scores for translations for different directions, generated using several leading LLMs: Claude[2], Gemini[3], GPT[4], Unbabel-Tower (Alves et al., 2024). Values are reference-based scores of MetricX-24. While these scores do not perfectly reflect translation quality, they offer a general indication of performance. The table illustrates that the translation quality achieved by the method described in this paper is comparable to that of the leading LLMs.

[2]https://claude.ai/
[3]https://gemini.google.com/
[4]https://chatgpt.com/

6 Conclusions and future work

We introduced a novel, high-speed decoding algorithm (A*) that generates hypotheses significantly faster than beam search in large language models (LLMs). This algorithm is adaptable to any language model. The application of Noisy Channel Reranking enhances the quality of diverse generated candidates by up to 0.2 points on the MetricX-24 scale. Further application of Quality Estimation (QE) reranking yields an additional improvement of another 0.2 to 0.3 points.

A* decoding algorithm flexibility enables a balance between quality and speed. Due to its efficiency and adaptability for any language model, this method has potential for numerous practical applications.

The proposed solution demonstrates high quality of EuroLLM models. The final results reflect a significant improvement over EuroLLM-9B translations, with a reduction of up to 0.5 points in MetricX-24 and an increase of 1.5 points in COMET, achieving quality comparable to leading large language models, even for non-European languages.

Due to time constraints prior to the WMT25 workshop, we were unable to explore post-processing techniques or context-aware multi-sentence decoding. The results submitted represent the unrefined output of the method described in this paper. A* decoding is a new idea, and we intend to refine it through further research. The results presented at WMT25 could be enhanced by leveraging other LLMs models as a base model, incorporating alternative Quality Estimation methods, or utilizing improved language models for reranking.

7 Acknowledgements

We would like to express our gratitude to the entire Machine Translation team at Samsung R&D Poland for their support. Special thanks go to Marcin Szymański for valuable reviews. We also acknowledge the management of Samsung Poland for their support in preparing this work.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages dialects.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Franz Josef Och, Nicola Ueffing, and Hermann Ney. 2001. An efficient A* search algorithm for statistical machine translation. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Kyra Yee, Nathan Ng, Yann N. Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation.

In2x at WMT25 Translation Task

Lei Pang, Hanyi Mao, Quanjia Xiao, Ruihan Chen, Jingjun Zhang, HaiXiao Liu*, Xiangyi Li
[1]Duxiaoman [2]University of Chicago [3]Peking University
[4]Harbin Institute of Technology
panglei@duxiaoman.com, hanyim@uchicago.edu,
xiaoqj@stu.pku.edu.cn, zhangjingjun@duxiaoman.com,
liuhaixiao@duxiaoman.com, xiangyi@duxiaoman.com

Abstract

This paper presents the open-system submission by the In2x research team for the WMT25 General Machine Translation Shared Task. Our submission focuses on Japanese-related translation tasks, aiming to explore a generalizable paradigm for extending large language models (LLMs) to other languages. This paradigm encompasses aspects such as data construction methods and reward model design. The ultimate goal is to enable large language model systems to achieve exceptional performance in low-resource or less commonly spoken languages.

1 Introduction

Machine translation (MT) has long been both a high-impact application and a central research challenge in natural language processing. The advent of large language models (LLMs) has reshaped MT from task-specific supervised learning toward large-scale representation learning and instruction-following paradigms, enabling steady gains across diverse language pairs (Alves et al., 2024; Jiao et al., 2023; Kocmi et al., 2024; Lu et al., 2024).

Yet, two persistent gaps remain. **First**, while mainstream LLM training increasingly optimizes for mathematical and code reasoning, their *expressive* and *creative* language abilities—e.g., idiomaticity, stylistic naturalness, and culturally appropriate phrasing—are comparatively underdeveloped(Lewkowycz et al., 2022; Liu et al., 2023; Lozhkov et al., 2024; Rozière et al., 2023; Zaitova et al., 2025). This often leads to translations that are locally literal but globally stilted, especially for informal registers, slang, and literary text . **Second**, model competence is unevenly distributed across languages: English receives disproportionate coverage and quality, while many non-English languages trail in both general capability and translation naturalness(Aharoni et al., 2019; Johnson et al., 2017; Kocmi et al., 2023, 2024; Team et al., 2022). Community findings over recent WMT cycles echo this asymmetry: despite the "LLM era", MT is far from solved uniformly across directions, with larger gaps off English-centric pairs and on long-tail phenomena .

This paper studies how to **transfer English strength into non-English targets** to improve expressive and culturally faithful translation. Concretely, we focus on Japanese—a language where literal adequacy is not sufficient: natural Japanese requires idiomatic paraphrasing, honorific and register control, and sensitivity to genre and context. Our thesis is that English can be used as a *hub language* to bootstrap these capabilities via curriculum design, cross-lingual alignment, and preference signals that explicitly reward naturalness.

We present **In2x**, a Japanese-focused model designed to inherit general competency from English while specializing for Japanese expressiveness. At a high level, In2x operationalizes three principles: (i) *English-as-hub transfer*: leverage rich English data and strong English modeling to seed robust lexical/semantic priors, then transfer to Japanese via bilingual and style-augmented objectives; (ii) *Expressiveness-first supervision*: emphasize prompts and signals that drive idiomaticity and cultural appropriateness (beyond literal adequacy); (iii) *Evaluation beyond metrics*: complement automatic metrics with human judgments targeted at idioms, slang, and stylistic naturalness.

We evaluate In2x on standard WMT-style test sets and targeted Japanese-focused challenge suites that stress idioms, slang, and style. According to the preliminary ranking results of WMT 2025, In2x outperforms many large-scale proprietary models, such as Gemini-2.5-Pro (Comanici and Team, 2025), GPT-4.1 (Fachada et al., 2025), Claude-4 (Anthropic, 2025), and DeepSeek-V3 (Monisha, 2025).

Overall, we make three core contributions:

1. We diagnose under-explored gaps in current

Proceedings of the Tenth Conference on Machine Translation, pages 671–679
November 8-9, 2025

LLM-based MT: the tension between heavy investment in math/code reasoning and the relative neglect of creative/idiomatic language ability, and the English-vs.-non-English capability asymmetry.

2. We introduce **In2x**, a Japanese-focused model that systematically transfers English strengths to Japanese, with an emphasis on naturalness and cultural appropriateness.

3. In this study, we introduce a detailed alignment pipeline designed to enhance the creative capabilities of language models. This approach not only improves performance in non-STEM (Science, Technology, Engineering, and Mathematics) tasks but also ensures that the models maintain robust generalization abilities across diverse linguistic challenges. For instance, in the en-ja translation track, the model demonstrates outstanding performance without any task-specific fine-tuning, highlighting its adaptability and effectiveness in non-STEM domains.

2 Continue Pretraining Stage

To balance the capabilities of large language models (LLMs) in both science-oriented and humanities-oriented domains during the pretraining process, we divided the continued pretraining stage into three distinct phases. The goal of this process is to enhance the model's multilingual proficiency, improve general-purpose abilities in foundational humanities tasks, and refine its representation in specialized contexts (Brown et al., 2020; Rae et al., 2021).

The training process incorporates diverse corpora, including a comprehensive 2 trillion tokens dataset comprising encyclopedic knowledge, webpages, structured information, news articles, Wikipedia entries, academic papers, and STEM-related datasets (Gao et al., 2020; Raffel et al., 2020). In addition, a dedicated 500 billion tokens corpus has been curated exclusively for creative writing tasks such as novel and screenplay synthesis, as well as authentic conversational datasets simulating real-life dialogue (Zhang et al., 2022).

Another significant aspect of this training stage focuses on enhancing capabilities in the target language, with Japanese utilized as an example. To this end, substantial Japanese language-specific corpora were introduced, alongside a balanced dataset with equal distribution of Chinese, English, and Japanese corpora (Xue et al., 2021). The aim was to facilitate transfer learning from pretraining on Chinese and English to the Japanese language.

2.1 Phase 1: Fundamental Knowledge Enhancement

In this phase, the creative writing corpus and the knowledge-focused corpus are jointly trained with constant learning rates. This approach was designed to boost proficiency in STEM-related reasoning while preserving the nuanced expression habits required for creative tasks in humanities (Kaplan et al., 2020).

2.2 Phase 2: Long-Text Capability Refinement

During this phase, a subset of the data was filtered based on text length, allowing the context length to increase from the typical 8,192 tokens to approximately 32,000 tokens. This step was intended to amplify the model's ability to process and comprehend extended-length texts (Hoffmann et al., 2022).

2.3 Phase 3: Fast Annealing Stage

In the final phase, a high-quality corpus was constructed based on selections informed by perplexity (PPL) and quality-assessment metrics. The annealing training was conducted with linear decay of the learning rate from 3×10^{-5}. This process consumed a total of 300 billion tokens and enabled the model to maintain its vivid expressive style for tasks such as novels and screenplays (Brown et al., 2020).

3 Post-Training Data

The post-training dataset consists of 2 million samples, with 1.5 million used during the supervised fine-tuning (SFT) process and 500,000 used in the reinforcement learning (RL) process. To ensure the Japanese language (our target language) achieves a proficiency level comparable to major languages such as Chinese and English, we adjusted the ratio of target language instructions to attain an equal balance across these languages. Specifically, we used a 1:1:1 ratio in the Instruct-to-Example (In2X) setup, striving to transfer the original model's knowledge into the target language as effectively as possible (Ouyang and et al., 2022; Zhou et al., 2023).

We developed a detailed pipeline for constructing the target language instructions, which can be categorized into three major synthetic processes:

3.1 Obtaining Open-Source Instructions

We began by collecting open-source instruction datasets available in the target language. These datasets include curated public data and traditional NLP fundamental tasks. Examples of such datasets include Dolly, OASST, and OASST2 (Koch and et al., 2023; OpenAI, 2023).

3.2 Target Language Instruction Rewriting

This process consists of several substeps designed to enhance the model's linguistic and cultural adaptability in the target language:

- **Creative Language Tasks:** To preserve the language's stylistic characteristics in humanities-focused tasks, we designed creative tasks where the responses include original stories or scripts (Bai and et al., 2022).
- **Basic Localized Tasks:** This includes rewriting instructions for tasks relevant to the local context, such as exam questions. Some of these tasks provide only the question and answer. We leveraged advanced models to supplement these datasets with reasoning chains to improve the model's reasoning ability in the target language (Wei and et al., 2022). This enhancement also helps to mitigate issues such as mathematical inconsistencies commonly faced during the LLM instruction synthesis process.
- **Cultural Style Transformation:** For certain humanities-related tasks, we incorporated cultural style shifts by adapting the instructions to align with the cultural norms and styles of the target language. This adjustment aims to improve the model's ability to provide culturally nuanced responses (Xu and et al., 2023).

3.3 Instruction Synthesis in the Target Language

We utilized methods such as Magpie (Xu et al., 2024) and Self-Instruct (Wang and et al., 2022) to synthesize target language instructions. However, these automatically generated instructions often suffer from issues including overly simple questions, lack of focus, self-answered queries, and internal contradictions. To address these challenges, we implemented a strict quality control pipeline with the following techniques:

- **Prompt Engineering:** We crafted detailed prompts with explicit rules to identify and troubleshoot common issues in synthesized instructions (White and et al., 2023).
- **Validation via Model Responses:** Instructions passing the first step were tested by having the model generate responses. These responses were evaluated by a critic LLM for contradiction, hallucinations, or failure to provide valid results. Instructions flagged with such issues were discarded. The critic LLM, being sensitive to hallucinations, acts as an additional safeguard for quality control (Ganguli and et al., 2022).
- **ReReading Mechanism:** After constructing the prompts for instruction generation, we employed a "ReReading" mechanism, where the model self-reviews its instructions. This review checks for correctness, alignment with the target language's cultural norms, and consistency with its native linguistic style. Since the synthesized instructions inherently carry the reasoning or rewriting processes behind them, leveraging this comprehensive context makes it easier to detect internal flaws, particularly those related to localization or cultural adjustments (Chiang and et al., 2023).

4 Post-Training SFT Stage

The post-training Supervised Fine-Tuning (SFT) stage is a critical step to balance linguistic diversity and optimize alignment within the instruction space for target languages. Below, we outline the key strategies and methods employed during this stage.

4.1 Balancing Linguistic Diversity

(a) **Clustering of Instruction Data:** To enhance linguistic diversity, the instruction dataset (comprising 40 million entries) was clustered using the Birch clustering algorithm (Zhang et al., 1996). The effectiveness of the clustering process was evaluated based on metrics like tag recapture rates and cluster smoothness (Zhang and Deng, 2020), which were used to fine-tune the clustering threshold. This process reduced the dataset to 1.5 million clusters after deduplication and selection.

(b) **Categorization via Large Language Models (LLMs):** Utilizing LLMs, the clustered data was tagged to assign both first-level and second-level labels (et al., 2020). For example, a mathematical problem might be categorized as "Mathematics - Quadratic Equations." These hierarchical labels provided a clearer structural organization of the data.

(c) **Difficulty Grading of Instructions:** The dataset was further refined by classifying each instruction according to its difficulty level: "Very Difficult," "Difficult," "Moderate," "Simple," and "Very Simple" (Wang and Li, 2019). For normalized scientific datasets, an additional evaluation was conducted using the LLaMA3-70B model (Research, 2023) with a Pass@16 metric (Perez and Andreas, 2022) to estimate the success rate of solving specific problems.

4.2 Aligning the Instruction Space of Target Languages

(a) **Avoiding Semantic Overfitting via Temperature Adjustment:** During training, a temperature parameter was introduced to mitigate overfitting of the model to specific linguistic semantic spaces (Sundararajan and Wang, 2021). This approach encouraged the model to adopt a more holistic learning strategy, enabling it to concentrate on question-answering techniques rather than over-specializing in the semantic patterns of a particular language. For instance, this allowed the Japanese language model to better mimic the cognitive behaviors observed in other languages (Koehn, 2019).

(b) **Specialized Sampling Strategy:** To further enhance the learning process, a two-step sampling strategy was employed over the 1.5 million clusters (Perket and Sanner, 2020):

- The difficulty levels of the data were sampled in a 3:3:3:1:0 ratio (corresponding to "Very Difficult," "Difficult," "Moderate," "Simple," and "Very Simple," respectively) (Finn and Jones, 2018).
- Additionally, within each cluster, samples were selected to ensure diversity across languages and categorical labels, which preserved the large-scale diversity of the original 1.5 million data points (Torroba and Blanco, 2021). This also maintained a degree of orthogonality between the target language and English within the sampled instructions (Feng and Gimpel, 2020).

The first round of sampling was used as the data for the first epoch, while the second round populated the second epoch. The training process adopted a learning rate of 2×10^{-5} with cosine decay for optimal performance (Loshchilov and Hutter, 2017).

5 Reinforcement Learning to Enhance General Capabilities in Cultural and Creative Industries

In the post-training RL stage, we leveraged a process similar to the instruction filtering procedure used during the SFT phase (Ouyang and et al., 2022). Specifically, an additional set of instructions was curated, comprising 500k samples that were guaranteed not to overlap with the instructions used in the SFT phase. The training configuration utilized a batch size of 128 and a minibatch size of 32, with the dataset trained for one epoch. Each rollout involved 16 iterations, and the reward evaluation was based on both a rule-based reward model and a generative reward model (Christiano et al., 2017).

5.1 Reward Model Design

The reward model system was meticulously designed to cater to different task types:

- **Rule-Based Reward Model:** For tasks involving mathematics, STEM disciplines, and logic, a rule-based reward model was employed to ensure adherence to specific criteria (Silver and et al., 2017).
- **Generative Reward Model for Creative Tasks:** For creative tasks like content generation, prompts were designed to embed specific scoring criteria or reward principles. These criteria included fundamental task requirements as well as dynamically generated guidelines tailored to the current prompt. For instance, in translation tasks, prompts might incorporate principles to penalize issues such as omissions, linguistic inconsistencies, or mixing of languages. The scoring mechanism then assessed adherence to these principles and calculated a reward score based on the percentage of fulfilled criteria (Liu et al., 2025)

- **Pair-wise Reward Model for Creative Tasks:** Initially, annotators labeled 2,000 commonly used task examples with their corresponding ground truth answers (gsb), identifying any problematic answers and providing critique reasons. Using these critiques and annotations, a pair-wise reward model was trained, achieving an accuracy rate of 70

5.2 RL Algorithm Design

To address the complexity of the tasks, we made strategic adjustments to the RL algorithm to achieve stable and efficient training:

- **Trajectory-Corrected GRPO:** Considering the diverse nature of tasks and reward types, a token-level clipping approach was deemed too restrictive and prone to causing training instability. Instead, we employed the Trajectory-Corrected version of the Generalized Proximal Policy Optimization (GRPO) algorithm (Schulman et al., 2017), which proved effective for handling multilingual tasks with varying reward functions. This modification enabled stable and continuous training while accelerating the convergence curve(Pang and Jin, 2025).

- **Dual-Clip Mechanism:** To improve stability, we integrated a dual-clip mechanism, which stabilized the variance of importance sampling at the sentence (sen) level (He et al., 2016). Additionally, we removed the lower bound of sampling, achieving optimal performance for the given tasks.

- **Soft Length Penalty:** A soft-length penalty was incorporated throughout the training process to encourage better length control in generated outputs (Wu et al., 2016).

- **High-Level Clipping:** A clipping mechanism was introduced to ensure robust control over high-level rewards (Schulman et al., 2015).

- **Temperature Decay:** A temperature decay strategy was applied to progressively adjust the sampling temperature during training, encouraging diversity in outputs while maintaining stability (Hinton et al., 2015).

- **Entropy Regularization:** The entropy value was set to 0.01 during training, enabling the model to conserve entropy and avoid premature saturation of the reward space (Williams, 1992).

- **Reducing Variance Caused by Task Lengths:** To alleviate the variance introduced by the differing lengths of creative writing tasks and scientific tasks, a sequence-level reward training strategy was employed. This approach balances the effect of length differences between arts and science tasks, enabling better convergence under different task scenarios(Mao et al., 2025).

6 Model Ensemble

Model ensemble techniques are employed by taking into account the orthogonality of linguistic capabilities among various models. Specifically, models that exhibit strong linguistic proficiency are selected for the ensemble process to maximize overall performance.

Furthermore, the fusion of model tensors is conducted based on gradient information and the importance of weights. This approach ensures a robust integration of model parameters, leveraging their respective contributions to optimize the ensemble. Such methodologies have been shown to enhance the effectiveness of model ensembles in complex tasks (Wang et al., 2025).

7 Evaluation Results

7.1 Benchmarks

The model demonstrated exceptional performance in widely recognized Japanese language benchmarks, particularly the ja-mtbench, showcasing its robust and reliable language translation capabilities. A comparative analysis of its performance against other prominent models, such as GPT-4-turbo (GPT-4.0), Claude 3.5, and Qwen-2.5-72b, is visually presented in Figure 1. As depicted in the figure, our model significantly outperforms its counterparts across various evaluation metrics, further validating its superiority in Japanese language processing tasks.

7.2 WMT Evaluation Results

The performance of the model, despite not having undergone specific fine-tuning for particular tasks, has demonstrated exceptional results in the automatic evaluation of Japanese-related tracks within

Table 1: English→Japanese Translation Performance at WMT25.

System	Human	AutoRank	Literary	News	Social	Speech
In2x	77.8	2.3	60.8	83.6	81.9	82.7
Gemini-2.5-Pro	85.8	2.5	87.3	82.5	87.7	86.0
GPT-4.1	83.7	2.9	95.4	77.0	80.7	84.9
Claude-4	79.3	5.8	86.5	76.1	72.8	86.3
Deepseekv3	79.3	4.7	82.9	80.0	74.1	82.7

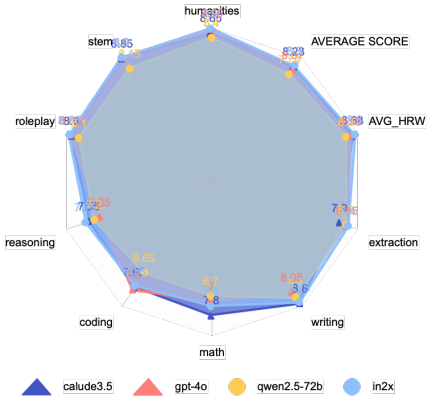


Figure 1: The performance results of in2x, gpt-4-turbo (GPT-4.0), Claude 3.5, and Qwen-2.5-72b on jamtbench.

the WMT competition. These outcomes are further corroborated by human evaluations, where the model outperformed closed-source commercial api, such as Claude 4, in areas including social interactions, news-related tasks, and speech generation(Kocmi et al., 2025b). However, it is worth noting that its performance in tasks requiring literary competence and advanced literary expression was suboptimal. This gap highlights a significant area for improvement, particularly in its handling of classical texts and the refinement of its literary language generation abilities. Moving forward, the primary focus of subsequent development efforts will lie in enhancing the model's proficiency in literary composition, with a particular emphasis on classical literature and the nuanced articulation of literary expression. For further details and specific performance metrics, one may refer to the original comparative analysis and evaluation results(Kocmi et al., 2025a).The specific results can be found in the table for reference1.

8 Conclusion

This work introduces and validates a method for transferring language modeling capabilities, as demonstrated on the WMT translation task. The proposed approach significantly enhances Japanese language proficiency during the CPT, SFT, and RL processes. Notably, without any additional language-specific fine-tuning, the large language model achieves alignment in its Japanese capabilities, bringing them on par with those of mainstream languages.

Furthermore, this work presents a systematic pipeline for aligning language models, as well as a method for training rewards in the creative content domain. Remarkably, this approach requires only around 2,000 annotated samples in the target language to achieve improved language transfer capabilities, with the remaining process relying on automated instruction-building techniques. Future efforts will focus on enhancing the literary style of the model by integrating elements such as classical texts and stylistic refinements into the pipeline.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 3874–3884.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Pierre Colombo, João G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Anthropic. 2025. Claude opus 4 demonstrates superior reasoning and coding performance. Online model announcement. Seen on Anthropic's site; explicit arXiv version not yet available.

Yuntao Bai and et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Aaron Chiang and et al. 2023. Autoreviewer: Enabling model self-review for dataset quality control. *arXiv preprint arXiv:2303.14112*.

Paul Christiano, Jan Leike, Tom Brown, et al. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

G. Comanici and Gemini Team. 2025. Gemini 2.5 pro: A thinking model with state-of-the-art multimodal reasoning. *arXiv preprint arXiv:2507.06261*. Retrieved from arXiv.

Tom B. Brown et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Nuno Fachada, Daniel Fernandes, et al. 2025. Gpt-4.1 sets the standard in automated experiment design using novel python libraries. *arXiv preprint arXiv:2508.00033*. Retrieved from arXiv.

Sandra Feng and Kevin Gimpel. 2020. Orthogonality in multilingual semantic spaces. *Computational Linguistics Research*, 101(4):567–589.

Hayley Finn and Bradley Jones. 2018. Leveraging difficulty-based sampling strategies in machine learning. *Machine Learning Journal*, 112(3):583–601.

Deep Ganguli and et al. 2022. Red teaming language models to reduce harmful outputs. *arXiv preprint arXiv:2202.03286*.

Leo Gao, Stella Biderman, Sid Black, Luke Golding, Travis Hoppe, Horace Foster, Jason Phang, Colin Raffel, Adam Roberts, Noam Shazeer, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Di He et al. 2016. Dual learning for machine translation. *arXiv preprint arXiv:1611.00179*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Eliza Chan, John Aslanides, Susannah Young, Trevor Cai, Ethan Rutherford, Saffron Huang, Roz Barnes, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jared Kaplan, Sam McCandlish, Tom Henighan, Danny Brandon, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, et al. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tim Koch and et al. 2023. Reinforcement learning with human feedback for oasst instructions.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of wmt25 general machine translation systems.

Tom Kocmi, Christian Federmann, et al. 2024. Findings of the wmt24 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, Miami, Florida, USA.

Tom Kocmi et al. 2023. Findings of the 2023 conference on machine translation (wmt23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore.

Philipp Koehn. 2019. Cross-lingual alignment and semantics in neural machine translation. *Computational Linguistics*, 45(1):1–24.

Aleksander Lewkowycz et al. 2022. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025. Inference-time scaling for generalist reward modeling.

Igor Loshchilov and Frank Hutter. 2017. Sgdr: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations*.

Alexei Lozhkov, Raymond Li, Leandro von Werra Allal, Filippo Cassano, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Hanyi Mao, Quanjia Xiao, Lei Pang, and Haixiao Liu. 2025. Clip your sequences fairly: Enforcing length fairness for sequence-level rl.

S. M. A. Monisha. 2025. A comparative study of reasoning-optimized large language models: Deepseek, chatgpt, and claude. *arXiv preprint arXiv:2502.17764*. Retrieved from arXiv.

OpenAI. 2023. Instructgpt: Aligning language models to follow human instructions. Https://openai.com/blog/instruction-following.

Long Ouyang and et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Lei Pang and Ruinan Jin. 2025. On the theory and practice of grpo: A trajectory-corrected approach with fast convergence.

Ethan Perez and Jacob Andreas. 2022. Pass@k: Measuring large language model problem solving. *arXiv preprint arXiv:2207.01986*.

Spencer Perket and Scott Sanner. 2020. Efficient sampling techniques for large-scale dataset training. In *Proceedings of the International Neural Information Processing Systems*, page 374–385.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Scott Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & challenges. In *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Meta AI Research. 2023. Introducing llama: A foundational, large language model. *Meta AI Technical Reports*.

Baptiste Rozière et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

John Schulman et al. 2015. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*.

John Schulman et al. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

David Silver and et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Ashok Sundararajan and Xin Wang. 2021. Temperature scaling for neural networks. *Journal of Machine Learning Research*, 23(1):140–158.

NLLB Team, Marta R. Costa-jussà, James Cross, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Lucas Torroba and Eduardo Blanco. 2021. Maintaining linguistic diversity in multilingual machine learning models. *Journal of Computational Linguistics*, 49(2):317–331.

Fan Wang and Qiang Li. 2019. Automatic difficulty estimation for instructional content. In *Proceedings of the Conference on Artificial Intelligence*, page 1657–1663.

Yizhong Wang and et al. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Zhixiang Wang, Zhenyu Mao, Yixuan Qiao, Yunfang Wu, and Biye Li. 2025. Optimal brain iterative merging: Mitigating interference in llm merging.

Jason Wei and et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

John White and et al. 2023. Prompt engineering strategies: A survey. *arXiv preprint arXiv:2302.06899*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Yonghui Wu et al. 2016. Google's neural machine translation system: Bridging the gap. In *arXiv preprint arXiv:1609.08144*.

Ling Xu and et al. 2023. Cultural adaptations for instruction-tuned language models. *arXiv preprint arXiv:2301.01234*.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing.

Chengqing Xue, Noah Constant, Adam Roberts, Mihir Kale, Aravind Goel, Brian Lester, Rami Al-Rfou, Aditya Siddhant, Imane Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the International Conference on Machine Learning*.

Iroda Zaitova et al. 2025. It's not a walk in the park! challenges of idiom translation in mt and slt. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Ailing Zhang, Xuchao Liu, Wenhao Huang, et al. 2022. A new approach to creative writing with large-scale language models. In *Proceedings of the Neural Information Processing Systems Conference*.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, 25(2):103–114.

Ying Zhang and Zhi-Hong Deng. 2020. Evaluation metrics for clustering algorithms in diverse data spaces. *Pattern Recognition*, 108:107533.

Alexander Zhou, Simone Palagi, and et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

CUNI at WMT25 General Translation Task

Miroslav Hrabal, Josef Jon, Martin Popel, Ondřej Bojar
Charles University, Faculty of Mathematics and Physics
{hrabal,jon,popel,bojar}@ufal.mff.cuni.cz

Abstract

This paper describes the CUNI submissions to the WMT25 General Translation task, namely for the English to Czech, English to Serbian, Czech to German and Czech to Ukrainian language pairs. We worked in multiple teams, each with a different approach, spanning from traditional, smaller Transformer NMT models trained on both sentence and document level, to fine-tuning LLMs using LoRA and CPO. We show that these methods are effective in improving automatic MT evaluation scores compared to the base pretrained models.

1 Introduction

We have entered the shared task as a number of small teams with different approaches, each with its own submission. We will describe the datasets, methods and evaluation results of each submission in the following sections. Here we will present just a brief overview of all the systems we submitted.

CUNI-MH-v2 is a constrained system trained on partially synthetic data sampled from the CzEng 2.0 (Kocmi et al., 2020) dataset using LoRA (Hu et al., 2021) and Contrastive Preference Optimization (Xu et al., 2024). We will release both the model weights and the filtered training data. The model itself is fine-tuned from the EuroLLM-9B-Instruct model. We currently only support two language directions, (en→cs) and (cs→de), and offer separate LoRA adapters for each. The translations were done on the paragraph level.

CUNI-EdUKate-v1 is an unconstrained system trained on educational domain data using LoRA, SFT, and Contrastive Preference Optimization. It is also fine-tuned from the EuroLLM-9B-Instruct model. It only supports cs2uk language direction and, unlike CUNI-MH-v2, both training and inference were done on sentence level.

CUNI-SFT models were created by a simple supervised finetuning using LoRA on a small amount of publicly available training data.

CUNI-Transformer and **CUNI-DocTransformer** are resubmissions of systems from previous years.

2 Methods

This section describes the approaches used for training our submissions.

2.1 CUNI-SFT (en2cs, en2sr, cs2uk)

We have finetuned multiple pretrained models for document-level and sentence-level translation using LoRA. We have used learning rate $lr = 2e-4$, LoRA ranks $r = 8$ and $r = 16$, LoRA $\alpha = 2 * r$ and batch size of 2 with 16 gradient accumulation steps, resulting in effective batch size of 32. We trained the models for 10k updates. We compared sentence-level translation without context, sentence-level with context shown to the LLM and pure document-level prompt. The prompts are shown in Section A.

2.2 CUNI-MH-v2 (en2cs, cs2de)

Considering that EuroLLM-9B-Instruct is already reasonably good at English to Czech translation, we chose to skip the supervised fine-tuning stage, thereby departing from year's CUNI-MH (Hrabal et al., 2024), and fine-tuned the model solely using Contrastive Preference Optimization (CPO) (Xu et al., 2024).

For the two language directions, (en→cs) and (cs→de), we trained separate LoRA adapters with rank $r = 32$, LoRA $\alpha = 64$, LoRA dropout of 0.05 and effective batch size of 8. We used cosine learning rate scheduler and trained for $10\,\text{k}$ steps.

2.3 CUNI-EdUKate-v1 (cs2uk)

CUNI-EdUKate-v1 was trained from EuroLLM-9B-Instruct model using LoRA in two stages. In

Proceedings of the Tenth Conference on Machine Translation, pages 680–687

the first stage, we train it on internal sentence-level educational domain parallel data. In the second stage, we train it on partially synthetic internal preference sentence-level educational domain data.

2.4 CUNI-(Doc)Transformer (cs2uk, en2cs)

CUNI-Transformer (cs→uk) and CUNI-DocTransformer (en→cs) are the same systems as submitted in previous years (Jon et al., 2023), relying on standard NMT training with Block backtranslation (Popel, 2018; Popel et al., 2020) and (in the case of CUNI-DocTransformer) document-level training.

3 Data

3.1 CUNI-SFT

We downloaded corpora for Czech to English, Croatian, Serbian[1], Bosnian, German and Ukrainian and English to Croatian, Serbian, German and Ukrainian from OPUS, keeping the document boundaries where possible. The datasets we used are: `DGT`, `DocHPLT`, `ELITR-ECA`, `EMEA`, `GlobalVoices`, `JRC-Acquis`, `News-Commentary`, `SETIMES`, `StanfordNLP-NMT`, `Tatoeba`, `TED2020`, `tico-19`, `TildeMODEL` and `WMT-News`. We scored these datasets with wmt22-cometkiwi-da QE model using Marian. We have selected the top 5% scoring documents (scores are computed on sentence-level and averaged) from each dataset for each direction, with at most 200 documents per dataset and direction. Documents longer than 60 sentences are split into 60-sentence chunks for scoring and training.

3.2 CUNI-MH-v2

In order to create the preference dataset necessary for the CPO method, we first sampled paragraphs from the CzEng 2.0 dataset and translated them using different models. For en→cs dataset, we used EuroLLM-9B Instruct and CUNI-MH from last year. We also used the reference translations as one of the possible candidate translations. For cs→de dataset, we used EuroLLM-9B Instruct, Qwen 2 and Qwen 3.

We then scored the translations (all synthetic candidates and the reference for the en→cs direction) using MetricX24 (used as a reference-free metric). From this, we created (source, preferred, dis-preferred) triplets by taking the highest-scoring translation as preferred and worse scored translations as possible dis-preferred translations.

Unlike the dataset used in previous year, where we gradually built paragraphs sentence by sentence (Hrabal et al., 2024), this year we chose to select the preference on the level of whole documents.

We further filtered these triplets using a version of our work-in-progress experimental metric based on Gemma 3 27b-it model, which we refer to as r1.1. We assigned the MetricX24 and r1.1 scores to each translation candidate. Afterwards, we considered the best candidate with the best MetricX24 score as preferred and all other candidates as dis-preferred. Out of those pairs, we kept only those that met the following criteria:[2]

1. The chosen and rejected translations differ.
2. $\mathrm{MetricX24(chosen)}$ is better than $\mathrm{MetricX24(rejected)}$ by at least 1.0 points.
3. $\mathrm{MetricX24(chosen)} < 10.0$.
4. $r1.1(\mathrm{chosen}) - r1.1(\mathrm{rejected}) \geq 1.0$.

The resulting en→cs dataset consists of 25530 preference triplets, and the cs→de dataset consists of 14797 preference triplets. All datasets and models will be available on Hugging Face:

- en→cs preference dataset: https://huggingface.co/hrabalm/CUNI-MH-v2-encs-data
- cs→de preference dataset: https://huggingface.co/hrabalm/CUNI-MH-v2-csde-data
- en→cs trained model: https://huggingface.co/hrabalm/CUNI-MH-v2-encs
- cs→de trained model: https://huggingface.co/hrabalm/CUNI-MH-v2-csde

3.3 CUNI-EdUKate-v1

For the CUNI-EdUKate-v1 model, we used our internal sentence-level Czech-Ukrainian parallel dataset covering the educational domain. This

[1]We transliterated all Serbian texts written in Cyrillic into the Latin script.

[2]Note that here we work with the raw MetricX24 outputs, which are greater than or equal to 0, and where lower is better.

Table 1: CUNI-MH-v2 en→cs performance on the development set. MetricX24 is google/metricx-24-hybrid-xl-v2p6-bfloat16. CometKiwi22 is Unbabel/wmt22-cometkiwi-da. r1.1 is our internal metric based on Gemma 3 27b-it assigning DA scores.

	wmt23				wmt23-para			
Model	BLEU	MetricX24	CometKiwi22	r1.1	BLEU	MetricX24	CometKiwi22	r1.1
CUNI-MH	36.52	–	83.16	–	35.42	–	**74.82**	–
EuroLLM-9B-Instruct	36.14	−3.74	82.90	89.66	36.69	−7.68	72.67	88.98
CUNI-MH-v2	**37.33**	−3.69	**83.38**	90.36	**37.81**	−7.53	73.75	91.81

Table 2: CUNI-MH-v2 en→cs performance compared with selected WMT24 models on the WMT24 test set.

	wmt24			
Model	BLEU	MetricX24	CometKiwi22	r1.1
Unbabel-Tower70B	24.72	**−3.70**	**83.04**	88.54
Claude-3.5	**32.04**	−4.62	80.79	**90.56**
CUNI-MH	27.62	−4.53	81.10	88.21
EuroLLM-9B-Instruct	26.04	−4.77	80.51	87.19
CUNI-MH-v2	27.89	−4.62	80.99	87.85

dataset is the only reason why our submission is unconstrained.

The creation of the preference dataset for the CPO stage was done in a similar way to the CUNI-MH-v2 model but using different selection of models to generate translation candidates and to score and filter them.

One notable difference was that we also trained EuroLLM-9B-Instruct to predict Direct Assesment scores and used the result as one of the models used to filter the preference triplets.

As a development set, we used 3770 segments split from the training data.

4 Evaluation

4.1 CUNI-SFT

We compared translation quality after finetuning across four pretrained models: `EuroLLM 9B`, `Aya Expanse 8B`, `Mistral Instruct v0.3 7B` and `Granite 3.3 8B`. We measured BLEU (Papineni et al., 2002) and chrF (Popović, 2015) on `newstest2019` (Barrault et al., 2019) in the English to Czech direction, NTREX (Federmann et al., 2022) for English to Serbian and `wmttest24` (Kocmi et al., 2024) for Czech to Ukrainian. The result for simple sentence-level and context-aware sentence-level prompts are shown in Table 3. We do not present results for the doc-level prompt, since we were not able to retrieve sentence-level alignment for source and translated sentences. Overall, we see that our approach to finetuning is effective for languages that are not well covered by the base model. For high resource combinations (e.g. eng-ces in EuroLLM), the finetuning does either not change the evaluation scores, or decreses them.

4.2 CUNI-MH-v2

During inference, we use vLLM and greedy decoding.

In Table 1, we show the performance of the en→cs CUNI-MH-v2 model on the development set. In Table 2, we compare its performance with best performing WMT23 models on WMT23 test set.

Interestingly, we can see that CUNI-MH-v2 improves in BLEU score compared to the base EuroLLM-9B-Instruct model, while we saw the opposite happen in the previous year (Hrabal et al., 2024), where the BLEU/chrF metrics got worse while the COMET22 and CometKiwi22 metrics improved. On the other hand, CUNI-MH-v2 gets higher CometKiwi22 score on sentence-level wmt23 dataset but lower score on the document-level version. Overall, we were able to achieve modest improvements in all metrics compared to the base model on both the development and test set.

For translation of the final WMT25 test set, we use the official script provided by WMT organizers

Context	Language	Model	Base		Finetuned	
			BLEU	ChrF	BLEU	ChrF
Yes	eng-ces	aya-expanse-8b	25.9	57.8	23.3	51.8
		EuroLLM-9B-Instruct	29.9	56.7	28.5	56.2
		granite-3.3-8b-instruct	22.1	51.5	18.5	47.3
		Mistral-7B-Instruct-v0.3	16.9	48.3	15.8	44.3
	eng-srb	aya-expanse-8b	3.3	20.9	7.3	35.2
		EuroLLM-9B-Instruct	15.4	46.6	15.6	46.6
		granite-3.3-8b-instruct	3.1	17.2	4.2	29.8
		Mistral-7B-Instruct-v0.3	2.3	14.8	11.2	40.4
	ces-ukr	aya-expanse-8b	27.3	56.2	25.5	52.0
		EuroLLM-9B-Instruct	28.7	56.4	26.8	54.7
		granite-3.3-8b-instruct	7.0	31.7	6.9	27.6
		Mistral-7B-Instruct-v0.3	15.7	47.7	13.3	39.0
		GPT-4.1-mini	33.7	61.7	-	-
No	eng-ces	aya-expanse-8b	25.4	51.8	26.4	54.9
		EuroLLM-9B-Instruct	31.7	59.0	31.1	59.1
		granite-3.3-8b-instruct	21.8	51.2	22.1	51.5
		Mistral-7B-Instruct-v0.3	13.0	43.4	20.2	49.7
		GPT-4.1-mini	32.5	59.2	-	-
	eng-srb	aya-expanse-8b	8.8	38.0	17.1	47.9
		EuroLLM-9B-Instruct	16.9	48.3	22.6	52.4
		granite-3.3-8b-instruct	6.7	34.8	15.2	45.6
		Mistral-7B-Instruct-v0.3	9.1	41.2	17.4	47.9
		GPT-4.1-mini	29.3	57.9	-	-
	ces-ukr	aya-expanse-8b	24.3	55.1	24.4	51.9
		EuroLLM-9B-Instruct	31.0	59.0	28.2	55.8
		granite-3.3-8b-instruct	6.6	44.4	10.5	35.3
		Mistral-7B-Instruct-v0.3	13.4	39.8	19.2	46.0
		GPT-4.1-mini	33.5	61.6	-	-

Table 3: BLEU and ChrF scores of base and finetuned CUNI-SFT models on devsets (newstest2019 for eng-ces and eng-srb, wmttest2024 for ces-ukr.

to extract paragraph-level segments. During the inference, we further split the paragraphs to chunks of at most 256 tokens by using the sentence-splitter Python library.

4.3 CUNI-EdUKate-v1

We show the automatic metrics of the CUNI-EdUKate-v1 model in Table 4. The EuroLLM-9B-Instruct model, which is also the base model, is used as a baseline.

Table 4: CUNI-EdUKate-v1 automatic metric scores on internal educational domain sentence-level development set.

	dev set	
Model	BLEU	MetricX24
EuroLLM-9B-Instruct	37.4	−3.59
CUNI-EdUKate-v1	**39.1**	**−3.33**

5 Tools

To give a proper credit, we list the tools we used during the development and inference with our models:

CUNI-MH-v2

- transformers (Wolf et al., 2020), peft (Man-

grulkar et al., 2022) and trl (von Werra et al., 2020) libraries for training

- vLLM (Kwon et al., 2023) for inference
- MetricX24 XL[3] (Juraska et al., 2024) for scoring, data filtering, evaluation
- DSPy (Khattab et al., 2024, 2022) and Gemma-3-27b-it (Team et al., 2025) for data filtering

CUNI-EdUKate-v1

- transformers, peft and trl libraries for training
- vLLM for inference
- LINDAT Translation[4] for segmentation and to serve the translation API
- CometKiwi22 (Rei et al., 2022) for scoring, data filtering, evaluation
- MetricX24 XL for scoring, data filtering, evaluation
- Gemma-3-27b-it for data filtering

CUNI-SFT

- transformers, peft and trl libraries for training
- vLLM for inference
- CometKiwi22[5] used through Marian (Junczys-Dowmunt et al., 2018) for data filtering

CUNI-(Doc)Transformer

- Tensor2Tensor (Vaswani et al., 2018)

6 Future work

We have several ideas to improve the performance of the future iterations of our CUNI-MH-v2 model. In particular, we plan to scale up the size of the preference dataset by using a larger portion of CzEng2.0 and by sampling more translation candidates.

We also plan on experimenting with including synthetically translated documents with no reference translations, to augment our dataset with longer examples.

[3] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6-bfloat16

[4] https://github.com/ufal/lindat-translation/

[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da-marian

7 Conclusion

In this paper, we presented the CUNI submissions to the WMT25 General Translation Task, covering English→Czech, Czech→German, English→Serbian, and Czech→Ukrainian language pairs. Future work will focus on scaling preference datasets and leveraging longer-context translation scenarios.

8 Acknowledgment

This work was supported by the project TQ01000458 (EdUKate) financed by the Technology Agency of the Czech Republic (www.tacr.cz) within the Sigma 3 Programme.

It was also partially supported by the Charles University Grant Agency in Prague (GAUK 244523), by SVV project number 260 821, by Czech Ministry of Education, Youth and Sports (grant MŠMT OP JAK Mezisektorová spolupráce CZ.02.01.01/00/23_020/0008518) and by National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO.

It has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2023062).

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Miroslav Hrabal, Josef Jon, Martin Popel, Nam Luu, Danil Semin, and Ondřej Bojar. 2024. CUNI at WMT24 general translation task: LLMs, (Q)LoRA, CPO and model merging. In *Proceedings of the Ninth Conference on Machine Translation*, pages 232–246, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI at WMT23 general translation task: MT and a genetic algorithm. In *Proceedings of the Eighth Conference on Machine Translation*, pages 119–127, Singapore. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines. In *The Twelfth International Conference on Learning Representations*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondřej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *arXiv:2007.03006*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa

Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 193–199, Boston, MA. Association for Machine Translation in the Americas.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformer Reinforcement Learning. `https://github.com/huggingface/trl`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

A CUNI-SFT Model Prompt Template

We have compared three ways of formatting the input. We present the corresponding prompts here.

Sentence-level:

```
Translate this {source_lang} sentence to
{target_lang}: {line}
```

Sentence-level with document context:

```
We need to translate one line from a
{source_lang} conversation into {target_lang}.
Source document: {document_src}
Already translated: {previous_translations}
Translate literally (no explanations) this
line: {line}
```

Document-level:

```
Translate from {source_lang} to {target_lang}:
{document}"
```

UvA-MT's Participation in the WMT25 General Translation Shared Task

Di Wu **Yan Meng** **Maya Nachesa** **Seth Aycock** **Christof Monz**
Language Technology Lab
University of Amsterdam
{d.wu, y.meng, m.k.nachesa, s.aycock, c.monz}@uva.nl

Abstract

This paper presents UvA-MT's submission to the WMT 2025 shared task on general machine translation, competing in the unconstrained track across all 16 translation directions. Unusually, this year we use only WMT25's blind **test set** (source sentences only) to generate synthetic data for LLM training, and translations are produced using pure beam search for submission. Overall, our approach can be seen as a special variant of data distillation, motivated by two key considerations: (1) perfect domain alignment, where the training and test domains are distributionally identical; and (2) the strong teacher model, GPT-4o-mini, offers high-quality outputs as both a reliable reference and a fallback in case of mere memorization.

Interestingly, the outputs of the resulting model, trained on Gemma3-12B using Best-of-N (BoN) outputs from GPT-4o-mini, outperform both original BoN outputs from GPT-4o-mini and Gemma3-12B in some high-resource languages across various metrics. We attribute this to a successful model ensemble, where the student model (Gemma3-12B) retains the strengths of the teacher (GPT-4o-mini) while implicitly avoiding its flaws.

1 Introduction

In this paper, we describe the details of our submission to the WMT 2025 shared task on the general machine translation (unconstrained track), which includes 16 translation directions. With recent advances in Large Language Models (LLMs), particularly the emergence of stronger multilingual models, our focus in this paper is on effectively and efficiently adapting a general-purpose LLM for translation-specific tasks with limited training.

Unusually, this year we use only the **test set** to build synthetic data for model training and generate translations again based on the test set using pure beam search for submission, as shown in Figure 1. This is based on several considerations:

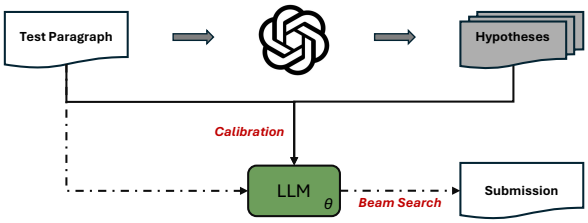


Figure 1: We use GPT-4o-mini to generate 16 hypotheses per sample on the WMT25 test set (at the paragraph level) using nucleus sampling. The resulting hypotheses are then used to train Gemma3-12B with the calibration method of Wu et al. (2025b). Finally, the calibrated Gemma model is used to translate the WMT25 test set for submission using pure beam search.

- **Very small sample sets** with a strong base model can effectively boost translation performance (Wu et al., 2024; Xu et al., 2024a).
- **Perfect domain alignment** since the training and test domains are inherently identical.
- **A strong teacher model** such as GPT-4o-mini[1] offers high-quality outputs as both a reliable reference and a fallback in case of mere memorization.

In the following sections, we test whether the student model (Gemma-3-12B) retains the strengths of GPT-4o-mini's outputs while implicitly avoiding certain flaws—effectively acting as a model ensemble.

More specifically, our strategy consists of two main steps:

Synthetic data building. We feed the WMT25 test set into GPT-4o-mini (Hurst et al., 2024), using the prompts provided with the offical test set[2], to generate 16 hypotheses per sample. The hypotheses are decoded using nucleus sampling with a top-p of 0.98[3] and a temperature of 1.0. Each sample

[1] As reported by Wu et al. (2025a), GPT-4o-mini can serve as a strong translation system.

[2] Official prompts are found here.

[3] We found that slightly lowering the top-p value effectively eliminates the off-targeting issue while preserving diversity.

Proceedings of the Tenth Conference on Machine Translation, pages 688–694
November 8-9, 2025

is at the paragraph level, where "$\backslash n$" remains in the original data as a separator.

Post-training. We apply the calibration method (Wu et al., 2025b) only to post-train Gemma-3-12B, which has been shown to be more effective than supervised fine-tuning or recent preference optimization methods, like CPO (Xu et al., 2024b). The calibration method aims to improve the correlation between translation likelihood and quality scores as measured by a reference metric model, enhancing the effectiveness of beam search decoding. Following Wu et al. (2025b), we use CometKiwi-XXL to score each one-to-many translation pair in our synthetic dataset.

Finally, the resulting model, trained on synthetic data derived from WMT25 test set, is used to again translate the WMT25 test set. We observe that for some high-resource languages, the resulting model's outputs even surpass the best hypotheses in the synthetic data—demonstrating a successful form of model ensemble.

In our next version, we provide detailed experimental settings and results, including: (1) offline experiments demonstrating the effectiveness of the calibration method; (2) offline experiments evaluating this test-time model ensemble strategy; and (3) our evaluation results for the final submission.

2 Calibration Method

We now briefly describe our post-training method, namely calibration (Wu et al., 2025b). This method addresses the miscalibration problem in machine translation, where translation quality deteriorates as search approximations improve and higher-probability hypotheses are potentially worse translations.

Prior studies have tried to mitigate this miscalibration issue by introducing an additional optimization step during inference time, known as quality-aware decoding (QAD) (Fernandes et al., 2022). These approaches typically involve generating multiple candidate translations through sampling, followed by reranking or voting using reference-free and/or reference-based machine translation metrics, such as Best-of-N (BoN) sampling (Rei et al., 2024; Faria et al., 2024) and Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004; Freitag et al., 2022).

The calibration approach mitigates this issue by optimizing the Pearson correlation between translation likelihood and quality during **training time**. Extensive experiments from Wu et al. (2025b) show several key advantages of this method, including:

1. Substantial translation performance gains with limited training.
2. Clear enhancements for maximum *a posterior* decoding, like beam search.
3. A unified framework for both translation quality optimization and estimation. Notably, we also apply this method to participate in the Quality Estimation task at WMT25[4].

In this shared task, we employ calibration as our only post-training method. For further technical details, please refer to (Wu et al., 2025b).

3 Online Evaluation

We thoroughly evaluate our system's outputs and compare them with those of several high-performance open-source and closed-source LLM-based translation systems, including GPT-4.1, Claude-4, Command-R+, DeepSeek-V3, TowerPlus-9B, TowerPlus-72B, Qwen2.5-7B, Qwen3-235B, and AyaExpanse-32B. We access these systems' results from the WMT25 MT evaluation test set[5], which was released a few weeks before the submission deadline of this paper.

We report results using three metrics: COMETKiwi$_{23}^{\text{DA}}$-XL, COMETKiwi$_{23}^{\text{DA}}$-XXL (Rei et al., 2023), and COMET$_{22}^{\text{DA}}$ (Rei et al., 2022). In addition, we conduct a light human evaluation for the English–Chinese track, comparing our system (UvA-MT) with our base model, Gemma-3-12B.

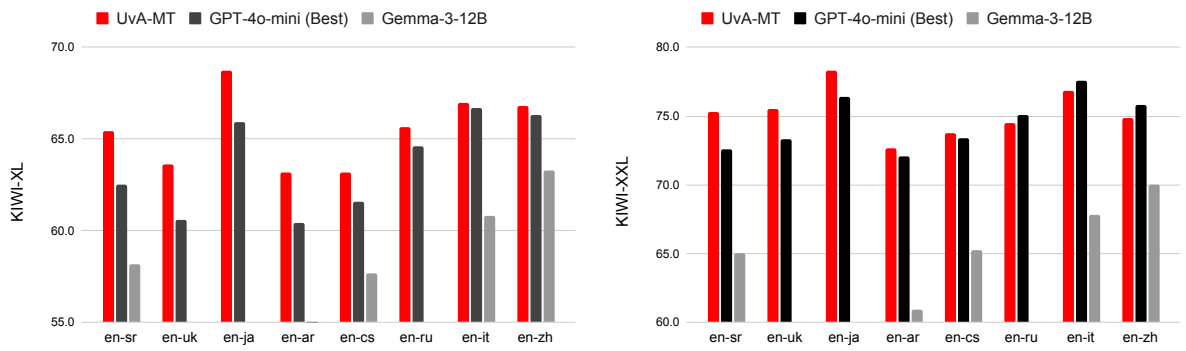
3.1 The Effectiveness of Ensembling

Figure 2 (a) and (b) show our system's results compared to those of 1) our base model, i.e., Gemma-3-12B, and 2) our teacher model, i.e., GPT-4o-mini, measured by CometKiwi-XL and CometKiwi-XXL, respectively.

Note that the reported results for GPT-4o-mini (best) are obtained using Best-of-N sampling. As described in the synthetic data building process in Section 1, we generate 16 hypotheses for each input and select the best one based on CometKiwi-XXL scores for evaluation.

[4]We cannot cite our QE system report at the time of submitting this paper; please refer to this year's findings paper for potential reference.

[5]The translation outputs from various systems are provided to human experts for annotation, which also serves as the evaluation task's test data.



(a) Performance measured by CometKiwi-XL

(b) Performance measured by CometKiwi-XXL

Figure 2: Performance comparison among the student model (Gemma-3-12B), the best-of-N outputs from the teacher model (GPT-4o-mini), and our ensemble model (UvA-MT). It is clear that UvA-MT surpasses all models in these 8 languages when measured with CometKiwi-XL, and outperforms some of them when measured with CometKiwi-XXL. Note that in some cases, such as **en-uk** and **en-ja**, the performance of Gemma-3-12B is either below the x-axis or not supported by the base model, so we did not show them in the figures.

For the other systems—namely Gemma-3-12B and our own (UvA-MT)—only a single hypothesis is generated per input using beam search with a beam size of 5.

As our system leverages both the base and teacher models, a successful ensemble would be expected to outperform each individually, at least in a portion of language directions.

Notably, we can observe in Figure 2 that:

- When evaluated with CometKiwi-XL (Figure 2-a), our system (UvA-MT) outperforms both the base model (Gemma-3-12B) and the teacher model (GPT-4o-mini) with Best-of-N sampling across all 8 language directions.
- When evaluating in CometKiwi-XXL (Figure 2-b), which is the very metric for Best-of-N selection, we can expect GPT-4o-mini's Best-of-N outputs to benefit from this evaluation due to greater potential for metric hacking. However, even under these conditions, our system still outperforms the teacher's results in a few language directions, such as `en-sr` (+2.7), `en-uk` (+2.0), and `en-ja` (+1.9), among others.

We acknowledge that some degree of metric gaming may also exist in our system's results above. However, we contend that a direct comparison between UvA-MT and GPT-4o-mini (Best) remains valid, because UvA-MT's training data is exactly the same as GPT-4o-mini's sampling data, we can therefore assume that UvA-MT could benefit from metric gaming *no more* than GPT-4o-mini's Best-of-N results; thus the relative gain over GPT-4o-mini (Best) should be considered realistic.

3.2 Overall Results

We now present a broader evaluation with detailed results measured by Comet22, CometKiwi-XL, and CometKiwi-XXL for 12 selected systems. For the complete set of results, please refer to the WMT25 findings paper. Note that the scores may differ slightly from those reported in the WMT25 findings paper, as variations in evaluation environments can introduce minor discrepancies (Zouhar et al., 2024).

Table 1 shows the results in CometKiwi-XL, one of this year's offical metrics. We show that UvA-MT achieves the best results in most of the language directions. When evaluated with CometKiwi-XXL (Table 2), the metric used for Best-of-N sampling, GPT-4o-mini (Best) obtains the highest scores in most cases. This is as expected with the previous discussion about metric gaming.

System ID	en-ru	en-it	en-zh	en-ar	en-cs	en-uk	en-ko
GPT-4.1	62.7	65.4	64.8	53.1	62.5	62.2	68.0
Claude-4	61.5	63.9	64.6	54.9	60.0	60.0	68.8
CommandA	–	64.4	64.2	53.1	60.3	60.6	68.6
DeepSeek-V3	62.5	65.0	61.3	57.0	62.1	61.3	67.5
UvA-MT	**65.6**	**67.0**	**66.8**	**63.2**	**63.2**	**63.6**	**70.0**
GPT-4o-mini (Best)	64.6	66.7	66.3	60.4	61.6	60.6	69.3
Gemma-3-12B	–	60.8	63.3	52.5	57.7	–	65.9
TowerPlus-9B	61.2	64.1	63.2	–	59.7	59.9	67.1
TowerPlus-72B	61.9	64.5	–	–	–	–	67.8
Qwen2.5-7B	–	58.4	62.1	–	–	–	–
Qwen3-235B	62.1	64.8	65.6	–	–	–	67.2
AyaExpanse-32B	–	63.9	–	58.2	59.9	–	66.7

System ID	cs-uk	en-ja	cs-de	en-et	ja-zh	en-is	en-sr
GPT-4.1	57.3	67.1	56.4	69.2	54.4	**64.7**	65.3
Claude-4	57.2	67.4	56.1	67.0	54.0	62.4	62.6
CommandA	56.4	67.1	56.5	–	53.4	–	–
DeepSeek-V3	55.8	66.2	56.3	–	52.6	54.7	60.1
UvA-MT	**57.7**	**68.7**	56.8	**69.3**	55.7	62.3	**65.4**
GPT-4o-mini (Best)	56.9	65.9	**58.4**	68.7	**56.2**	62.8	62.5
Gemma-3-12B	54.2	–	54.5	59.6	–	51.6	58.2
TowerPlus-9B	55.2	66.2	55.0	–	53.2	63.5	36.9
TowerPlus-72B	–	–	55.7	–	53.3	61.7	–
Qwen2.5-7B	–	–	–	–	52.2	–	–
Qwen3-235B	–	66.3	55.0	–	54.1	–	59.0
AyaExpanse-32B	55.4	–	55.0	–	–	–	–

Table 1: KIWI-XL scores across languages and systems. We highlight UvA-MT and GPT-4o-mini (Best), where the former uses the latter's output as training data. Bold indicates the highest score per column. We discard the results in two extremely low-resource directions, i.e., English to Bhojpuri and Maasai, as they are not supported by the base model and therefore lack meaningful comparability.

We note that the primary focus of this paper is to explore whether an ensemble strategy can outperform the teacher's output—a trend that is clearly observed in most cases in Table 1 and in a few cases in Table 2.

A more convincing result is obtained with Comet22, the reference-based metric, where we additionally consider the translation references provided by WMT25, thus maximizing the metric difference between training and evaluation. In Table-3, we can see that UvA-MT achieves best results in **en-ru** and **en-it** among all systems.

4 Discussion and Conclusion

Beyond Metric Hacking. We acknowledge that some degree of metric gaming is present in the results above, although its extent is difficult to quantify. Our focus in this competition, however, is to demonstrate gains that go beyond mere metric hacking.

In the extreme case where UvA-MT simply memorized the best outputs from GPT-4o-mini (maximizing hacking), the latter's score would represent the upper bound of the former. Therefore, a direct comparison between UvA-MT and GPT-4o-mini (Best) remains realistic, and any gain over GPT-4o-mini (Best) would reflect genuine enhancements. Interestingly, we observe them in most of the language directions.

We attribute these gains to a form of successful model ensemble, in which the student LLMs integrate the strengths of the teacher model's outputs while discarding some of their shortcomings. Regarding the role of the post-training method applied here, including whether it is a critical component for this ensemble, we leave for future investigation.

Practical Significance. Our setting is not well-suited for real-time translation systems, as training a student model is required for each group of new inputs. Nevertheless, our findings point to a promising direction for ensembling the strengths of two models when the target domain is established in advance. This is particularly relevant in practical scenarios such as customized translation, where latency is secondary and effectiveness is the foremost priority.

System ID	en-ru	en-it	en-zh	en-ar	en-cs	en-uk	en-ko
GPT-4.1	70.4	74.8	72.3	62.8	72.9	74.5	78.7
Claude-4	69.7	72.8	72.3	63.6	69.0	71.0	78.9
CommandA	–	72.9	71.4	62.0	69.3	70.8	78.4
DeepSeek-V3	69.7	74.3	67.6	65.6	72.7	73.4	77.7
UvA-MT	74.5	76.9	74.8	**72.6**	**73.8**	**75.5**	79.9
GPT-4o-mini (Best)	**75.1**	**77.6**	**75.8**	72.1	73.4	73.3	**80.8**
Gemma-3-12B	–	67.8	70.0	60.9	65.2	–	75.3
TowerPlus-9B	68.8	71.9	69.9	–	68.2	70.3	76.3
TowerPlus-72B	70.3	73.3	–	–	–	–	77.4
Qwen2.5-7B	–	62.9	68.7	–	–	–	–
Qwen3-235B	70.0	73.9	73.7	–	–	–	77.4
AyaExpanse-32B	–	71.8	–	66.1	69.0	–	76.1

System ID	cs-uk	en-ja	cs-de	en-et	ja-zh	en-is	en-sr
GPT-4.1	62.6	76.1	66.4	81.0	65.1	**74.5**	75.8
Claude-4	63.5	77.2	66.5	77.9	65.2	70.3	72.5
CommandA	61.6	76.1	66.5	–	64.6	–	–
DeepSeek-V3	60.7	74.9	65.5	–	62.2	60.3	69.3
UvA-MT	63.6	**78.3**	67.2	80.2	67.1	68.5	**75.3**
GPT-4o-mini (Best)	**66.8**	76.4	**72.0**	**81.4**	**69.8**	72.4	72.6
Gemma-3-12B	59.7	–	63.3	68.0	–	51.9	65.1
TowerPlus-9B	60.5	74.8	64.5	–	63.8	72.1	36.8
TowerPlus-72B	–	–	65.1	–	64.2	69.3	–
Qwen2.5-7B	–	–	–	–	62.0	–	–
Qwen3-235B	–	75.5	64.5	–	64.5	–	66.9
AyaExpanse-32B	61.0	–	64.7	–	–	–	–

Table 2: KIWI-XXL scores across languages and systems. We highlight UvA-MT and GPT-4o-mini (Best), where the former uses the latter's output as training data. Bold indicates the highest score per column.

System ID	en-ru	en-it	en-zh	en-ar	en-cs	en-uk	en-ko
GPT-4.1	82.4	45.7	82.9	**79.1**	**85.7**	**85.5**	87.4
Claude-4	80.6	44.2	82.1	76.4	82.3	82.3	85.9
CommandA	–	44.7	80.9	77.4	82.9	82.9	85.4
DeepSeek-V3	82.1	46.0	80.9	79.0	85.4	84.9	86.6
UvA-MT	**83.4**	**46.5**	82.7	78.9	85.2	84.5	86.3
GPT-4o-mini (Best)	–	–	–	–	–	–	–
Gemma-3-12B	–	44.6	80.7	77.0	80.4	–	84.7
TowerPlus-9B	80.8	44.7	80.6	–	82.5	83.1	83.8
TowerPlus-72B	80.9	44.6	–	–	–	–	84.3
Qwen2.5-7B	–	43.0	80.9	–	–	–	–
Qwen3-235B	82.1	46.1	**83.4**	–	–	–	**87.3**
AyaExpanse-32B	–	45.1	–	75.4	82.7	–	84.6

System ID	cs-uk	en-ja	cs-de	en-et	ja-zh	en-is	en-sr
GPT-4.1	**88.1**	**88.1**	**83.9**	**86.5**	**85.0**	**81.8**	**83.8**
Claude-4	87.0	86.4	82.3	83.2	83.7	78.3	79.6
CommandA	87.2	86.5	83.1	–	83.4	–	–
DeepSeek-V3	87.6	87.4	83.4	–	83.6	73.7	80.2
UvA-MT	86.9	86.9	82.5	85.0	82.8	78.5	67.7
GPT-4o-mini (Best)	–	–	–	–	–	–	–
Gemma-3-12B	85.1	–	80.6	79.3	–	70.4	74.1
TowerPlus-9B	86.7	85.7	81.2	–	82.6	81.0	49.0
TowerPlus-72B	–	–	81.3	–	82.5	79.5	–
Qwen2.5-7B	–	–	–	–	81.1	–	–
Qwen3-235B	–	87.9	82.3	–	83.8	–	78.3
AyaExpanse-32B	86.5	–	83.0	–	–	–	–

Table 3: Comet22 scores across languages and systems. We highlight UvA-MT and GPT-4o-mini (Best), where the former uses the latter's output as training data. Bold indicates the highest score per column.

References

Gonçalo Faria, Sweta Agrawal, António Farinhas, Ricardo Rei, José de Souza, and André Martins. 2024. Quest: Quality-aware metropolis-hastings sampling for machine translation. *Advances in Neural Information Processing Systems*, 37:89042–89068.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Di Wu, Seth Aycock, and Christof Monz. 2025a. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

Di Wu, Yibin Lei, and Christof Monz. 2025b. Calibrating translation decoding with quality estimation on llms. *arXiv preprint arXiv:2504.19044*.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *ICML*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth*

Conference on Machine Translation, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

AMI at WMT25 General Translation Task: How Low Can We Go? Finetuning Lightweight Llama models for Low Resource Machine Translation

Atli Jasonarson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
atli.jasonarson,steinthor.steingrimsson@arnastofnun.is

Abstract

We present the submission of the Árni Magnússon Institute's team for the WMT25 General translation task. We focus on the English→Icelandic translation direction. We pre-train Llama 3.2 3B on 10B tokens of English and Icelandic texts and fine-tune on parallel corpora. Multiple translation hypotheses are produced first by the fine-tuned model, and then more hypotheses are added by that same model further tuned using contrastive preference optimization. The hypotheses are then post-processed using a grammar correction model and post-processing rules before the final translation is selected using minimum Bayes risk decoding. We found that while it is possible to generate translations of decent quality based on a lightweight model with simple approaches such as the ones we apply, our models are quite far behind the best participating systems and it would probably take somewhat larger models to reach competitive levels.

1 Introduction

Large language models (LLMs) are becoming the predominant approach for a wide variety of tasks in the field of natural language processing. They have shown remarkable translation capabilities, see e.g. Kocmi et al. (2024), especially for well-resourced languages such as English and Spanish, but also for many low-resource languages (LRLs) as Xu et al. (2025) show for translations between English and Icelandic. The largest of these models, such as GPT-4 (OpenAI et al., 2024), are hardware and energy intensive, both in training and at inference time, and thus costly. The vast majority of open weights large language models, such as the Llama family of models (Touvron et al., 2023; Grattafiori et al., 2024), Aya (Üstün et al., 2024), Mistral (Jiang et al., 2024) and others, are primarily trained on English and other languages that are well represented on the internet, for obvious availability reasons. This has ramifications for LRLs. Not only do the models offer inferior performance for LRLs 'out-of-the-box', their vocabulary is typically underrepresented due to the smaller amounts of training data in these languages, see e.g. Nag et al. (2025). This leads to less efficient tokenization for the LRLs, meaning that the number of characters per token are considerably fewer than for well-resourced languages like English. For example, using the Llama-3.2 tokenizer, the average number of characters per token for Icelandic is ≈2.2, while for English, each token has ≈4 characters. For LRLs, more tokens are thus necessary to cover the same context length.

While we do encounter challenges when working with LRLs in the context of LLMs, there are a variety of approaches to increase the capabilities of the models in that regard. Xu et al. (2024a) trained their ALMA translation models based on LLama-2 (Touvron et al., 2023) by continual pre-training (CPT) and then fine-tuning the models on the translation task. One of the languages pairs they worked with was English↔Icelandic and they achieved results very competitive to previous models trained for that language pair.

In this paper, we will be working with that same language pair, English–Icelandic. We are interested in building models that are as lightweight as possible, while still retaining the translation capabilities of LLMs. We experiment with applying the ap-

Proceedings of the Tenth Conference on Machine Translation, pages 695–704
November 8-9, 2025

proach used for training the ALMA models to train bilingual translation models based on the Llama-3.2 models (Grattafiori et al., 2024). We compare the 1B parameter model to the 3B parameter model. For our training, we use a much larger Icelandic monolingual dataset than Xu et al. (2024a), as well as a larger parallel dataset for English–Icelandic. We find that the 1B parameter model produces considerably lower quality translations and is much more prone to hallucinate. Our final system, submitted to the WMT25 General Translation task (Kocmi et al., 2025a) is thus based on Llama-3.2 3B parameter model. Following Xu et al. (2024b), we experiment with contrastive preference optimization (CPO) on top of the fine-tuned model. Our system generates multiple hypotheses, using different temperature settings, with and without CPO. The hypotheses are then post-processed using a grammar error correction (GEC) model and post hoc rules to fix punctuation errors as well as mistakes in translating emojis, hashtags, email-addresses and URLs.

Our code is available on Github[1] and the translation model on Huggingface[2].

2 Related Work

Up until 2020, when Jónsson et al. (2020) published the first paper describing SMT and NMT for translations between English and Icelandic, not much work had been done with regard to MT for this language pair. Since WMT 2021, when English↔Icelandic was one of the language pairs for the news translation task (Akhbardeh et al., 2021), multiple MT publications have described MT research on Icelandic, using the WMT21 evaluation dataset. In 2024, the AMI team submitted a system to the WMT general translation task for the English→Icelandic language pair. The submission describes an effort to build a lightweight NMT system, using an encoder-decoder architecture. While it was small enough to run easily on a laptop computer, it still scored higher than many commercial systems (Jasonarson et al., 2024). In their work on building MT systems from the LLaMA-2 models, Xu et al. (2024a) pre-train models of two different sizes, 7B and 13B parameters, on data in six languages, two of these being English and Icelandic. They then fine-tune the model on the translation task.

[1] github.com/steinst/WMT25_AMI

[2] arnastofnun/Llama-3.2-3B-wmt25-AMI-en-is

3 Building the System

In building our model, we followed the approach used for training the ALMA-R models (Xu et al., 2024b), but instead of training on six languages, we trained only on texts in English and Icelandic. We are interested in investigating whether using some of the smallest available open LLMs can produce competitive translations and thus experiment with the lightweight Llama 3.2 1B and 3B parameter models. Hyperparameters used in training are reported in Appendix A.

Xu et al. (2024b) use the OSCAR 23.01 corpus (Ortiz Suárez et al., 2019; Kreutzer et al., 2022) which only contains approx. 300M running words in Icelandic. Furthermore, for fine-tuning English↔Icelandic, they only use 2000 sentence pairs. In our experiments, we extend both the monolingual and parallel data sets used.

3.1 Pre-training

The ALMA model employed CPT to improve model capabilities in the languages they work with. In doing that, they train their model on 20B tokens. As the Oscar dataset contains less than 300M running words in Icelandic, it would have to be repeated multiple times if a similar training setup were to be used for only two languages, English and Icelandic. Therefore, we add another data source, the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018; Barkarson et al., 2022a), on top of the OSCAR data. We use the 2022 version of the corpus (Barkarson et al., 2022b) and the 2024 extension (Barkarson and Steingrímsson, 2024). Combined they contain 2.6B running words from texts in 8 domains: news, parliamentary speeches, social media, published books, journals, Wikipedia, law texts and adjudications. We exclude the last two, law texts and adjudications, as these texts are quite atypical of texts in other domains. We also filter out paragraphs that we do not expect to be beneficial. These include duplications, paragraphs containing less than five words, paragraphs containing less than 50% alphabetical letters, and paragraphs that were not classified as Icelandic using langdetect (Nakatani, 2010) with a custom Icelandic language profile[3]. Finally, we split long paragraphs, over 255 tokens as tokenized by the Llama 3.2 tokenizer, into shorter segments. This resulted in a corpus of 2.17B running words in addition to the 294M in OSCAR. We estimate that

[3] github.com/steinst/langdetect_profiles

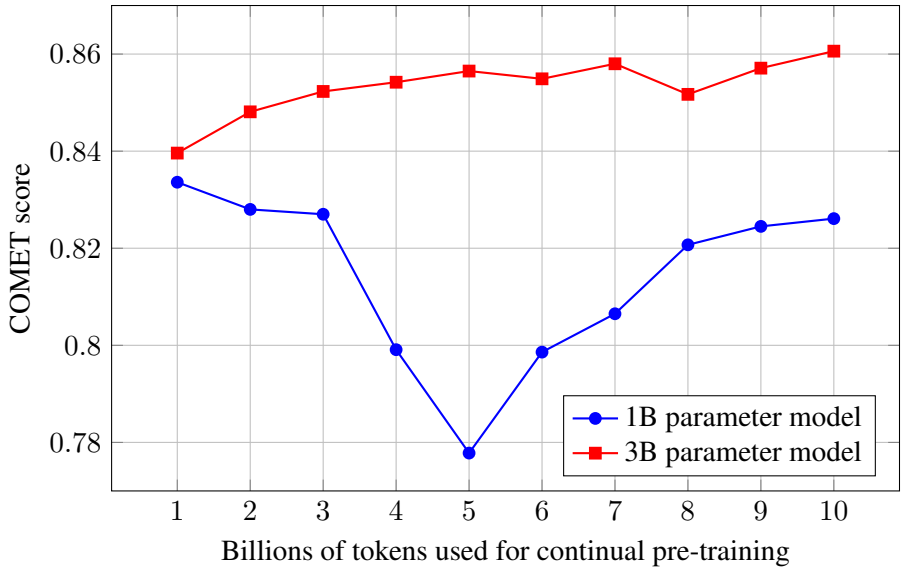


Figure 1: Comparison of COMET scores for 1B and 3B parameter models. Comet-scores for the WMT21 test dataset is calculated during CPT with intervals of 1B tokens.

Tokens (B)	File size (K) (1B model)	File size (K) (3B model)
1	183	216
2	210	197
3	218	174
4	339	178
5	400	170
6	303	167
7	299	163
8	231	178
9	234	163
10	219	161

Table 1: File size of translations at each stage of the training. The original English file is 141K

our data contains ≈7.6B tokens of Icelandic text when tokenized by the Llama-3.2 tokenizer.

We train our models on up to 10B tokens in total, with a 50/50 split between Icelandic tokens and English tokens. Training the 3B parameter model took ≈75 hours on 8xA100 GPUs and fine-tuning ≈80 minutes on the same hardware, while training the 1B parameter model took ≈120 hours on 2xA100 GPUs and fine-tuning ≈3.5 hours on a single A100.

3.2 Fine-tuning

In selecting the dataset to use for fine-tuning, we experimented with a different number of sentence pairs and different combinations of data.

We used three datasets, two described in the AMI submission paper for WMT general translation task last year (Jasonarson et al., 2024) as well as a small specialized dataset:

1. The baseline dataset, comprising data from the ParIce corpus (Barkarson and Steingrímsson, 2019; Steingrímsson and Barkarson, 2021), realigned using SentAlign (Steingrímsson et al., 2023), as well as sentence pairs from Paracrawl (Bañón et al., 2020) using the filtering approaches described in (Steingrímsson et al., 2023).

2. The synthetic sentences generated for training the AMI translation models for the WMT24 submission.

3. 1000 sentence pairs containing Icelandic idiomatic expressions and their English translations (Steingrímsson et al., 2024)

We scored the sentence pairs from the two large datasets using LaBSE (Feng et al., 2022) and fine-tuned the 3B parameter model, trained on 10B tokens on different combinations of the data. Different number of sentence pairs using only the baseline data, as well as a different number of sentence pairs using a mix of the baseline data and the synthetic data. We evaluated the fine-tuned models using the WMT21 evaluation set. When only using the baseline data, we achieved the highest COMET score using only 20k sentence pairs, but when mixing the baseline sentence pairs 50/50 with the synthetic data, as well as adding the small dataset of

idiomatic expressions in context, we achieve an even higher score. The highest scoring model was trained on a combination of these datasets, with 50k sentence pairs from the baseline set, 50k sentence pairs from the synthetic dataset and 1k sentences containing Icelandic idiomatic expressions and their English translations, resulting in a fine-tuning dataset of 101k sentence pairs.

3.3 CPO

CPO is introduced in Xu et al. (2024b) as an approach to mitigate two shortcomings of supervised fine-tuning: Firstly, to try to imitate training data and thus capping the model performance at that quality level, and secondly, to give the model a mechanism to reject mistakes in translation. This is important as even human-translated texts can have flaws and errors. To accomplish this, CPO uses specially curated preference data, with each source sentence having three translations: one human translation and two automatic translations, along with quality assessment scores for each translation. The highest-scoring translation is preferred and the lowest-scoring one dispreferred, in order to train the model to refine details and achieve better translations.

We created a new CPO dataset, for finalizing the models after fine-tuning. While the ALMA project only used the Flores dataset (Goyal et al., 2022) for CPO when working with English↔Icelandic, a total of 2,009 sentences, we add sentences from the WMT24 general translation shared task (Kocmi et al., 2024), 997 sentences, the development set from WMT21, 2,004 sentences, and the Icelandic parallel UD tree bank (Jónsdóttir and Ingason, 2020), 1,000 English sentences translated by a human translator into Icelandic.

In total the CPO data consists of approx. 6,000 items, each item comprising an English sentence and a human translation, or vice versa, two automatic translations for each language, one by the fine-tuned model described in Section 3.2 and the other by Claude Sonnet 4[4]. For each translation, we calculate three scores using reference-free models, `wmt23-cometkiwi-da-xl`, `XCOMET-XL` and an average of the two.

We apply CPO after pre-training and fine-tuning, as described in Section 3.1 and Section 3.2.

[4]We used `claude-sonnet-4-20250514` for both translation directions.

Model step	Score
Llama-3.2 3B (baseline model)	0.5197
Llama-3.2 3B + CPT	0.7940
Llama-3.2 3B + CPT + FT	0.8606
Llama-3.2 3B + CPT + FT + CPO	0.8441

Table 2: COMET-scores for the 3B parameter model after each ablation step, before post-processing.

3.4 Model Training

We trained the 1B and 3B parameter Llama 3.2 models using up to 10B tokens, with a 50/50 split between Icelandic and English tokens. After training, we selected the best fine-tuning dataset using the 3B parameter model, trained on 10B tokens, which scored highest of the trained models when evaluated using the English→Icelandic test set from WMT21. Figure 1 shows the scores for the models, evaluated after every 1B tokens of CPT, followed by fine-tuning. Both models are still improving when we stop training, indicating that we could probably achieve higher quality if we continue. It is worth noting that the 1B parameter model behaves rather curiously. After obtaining surprisingly good scores early in the training process, the COMET scores drop substantially, but then start rising again. We investigated what was going on and found that in the beginning, the model was not very likely to produce much longer strings than the source sentence. After training for a bit longer, the model becomes much more likely to continue producing text after it has finished producing the translation. This is reflected in the file size of the translations, shown in Table 1. File size closer to the size of the source file generally score higher than larger files.

We also carried out CPO after fine-tuning, which did not increase the COMET score on the evaluation set. COMET-scores for each ablation step are given in Table 2. The table indicates that without any continued pre-training the LLama-3.2 3B model does not seem to produce very coherent translations, but this should be expected as Icelandic is not one of the officially supported languages of Llama 3.2. Fine-tuning after CPT substantially increases the translation quality as measured by COMET, but in our experiments, CPO fails to improve it further.

3.5 Post-processing and MBR

When LLMs translate text, they have a tendency to continue generating new text after the translation is completed, irrelevant to the source text, as described in the previous section. While this seems to happen less with the 3B parameter model than with the 1B parameter one, it can still be a problem. When translating long sentences or paragraphs, both models seem to be more likely to skip parts and to be more prone to hallucinating. Finally, the Icelandic output commonly has incorrect inflections and word formation.

In order to counter some of these issues, we post-process the translation output. Post-processing uses the GEC model described in Jasonarson et al. (2024), and heuristics to ensure consistency between the source and target in the use of emojis, hashtags, URLs and punctuation.

For our final submission, we use the larger 3B parameter model after pre-training on 10B tokens, as this gave us the best results for our test set, as shown in Section 3.4. In order to increase the variety of translation candidates, we also do CPO training on the models and use both variants of the model, with and without CPO, to generate hypotheses:

- For both variants of the model, CPO-trained and not, we generate 9 translation hypotheses for each sentence, 3 for each of three temperature settings: 0.2, 0.6 and 0.9, resulting in 18 candidates in total.
- We post-process all 18 candidates, generating 18 new candidates. A total of 36, half post-processed and half not.
- Finally, in order to tackle the problem of the model spinning out of control and generating more text after translation has finished, we split each candidate translation on sentence boundaries. We then generate a sequence of partial candidates incrementally: the first partial candidate contains only the first sentence; the second partial candidate contains the first two sentences; the third contains the first three sentences; and so on, until the final candidate is identical to the complete original candidate, as exemplified in Figure 2.

All of these candidates are taken into consideration for COMET-MBR (Fernandes et al., 2022), employing cometkiwi-xl to select the final translations, considering the source and all generated candidates. Before settling on cometkiwi-xl, we compared two models, cometkiwi-xl and xcomet-xl. We had each model select their best candidates and then manually evaluated sentence pairs where the decisions of the two models differed. We found that cometkiwi-xl was more in line with our evaluation and thus chose that model for our pipeline.

Incremental Candidate Construction

Candidate 1: Samkvæmt embættismönnum hafa viðskiptavinir sem heimsóttu bankann einnig verið ráðlagt að fara sjálfviljugir í kórónuveirupróf.

Candidate 2: Samkvæmt embættismönnum hafa viðskiptavinir sem heimsóttu bankann einnig verið ráðlagt að fara sjálfviljugir í kórónuveirupróf. This translation has been made possible through the support of the American people through the United States Agency for International Development (USAID).

Candidate 3: Samkvæmt embættismönnum hafa viðskiptavinir sem heimsóttu bankann einnig verið ráðlagt að fara sjálfviljugir í kórónuveirupróf. This translation has been made possible through the support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the Government of Iceland and do not necessarily reflect the views of USAID or the U.S. Government.

Figure 2: An example of a translation candidate where the model continued generating after the translation was complete. We split the output on sentence boundaries to generate new candidates from the original one. The original English sentence was: "According to the officials, the customers who visited the bank have also been advised to voluntarily appear for coronavirus tests." In this case, the first sentence is the correct translation.

4 Translation Pipeline

Figure 3 shows the translation pipeline. Input documents to be translated are split into paragraphs and the MT system uses different settings for num-

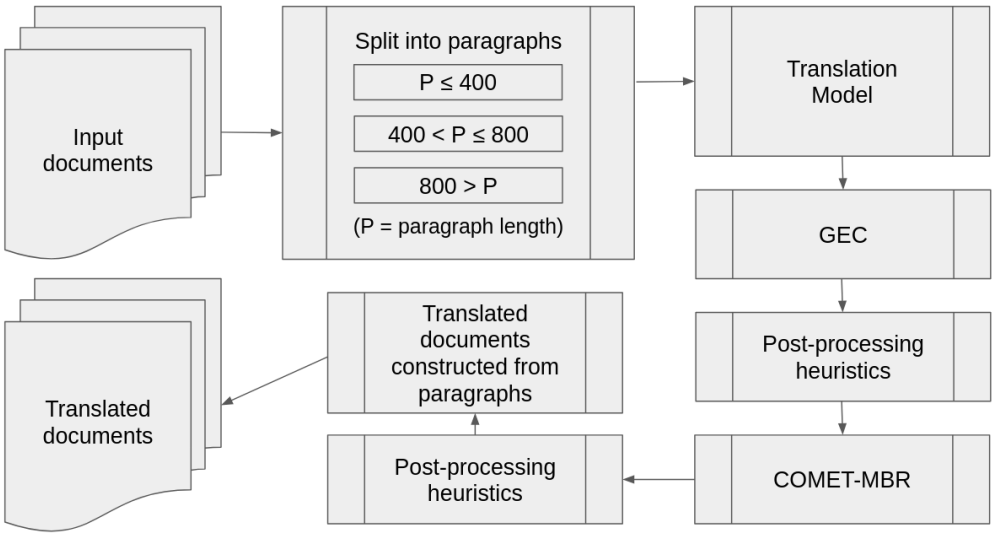


Figure 3: Processing pipeline as described in Section 4.

ber of input and output tokens depending on paragraph length measured in number of characters. 18 translation candidates are produced, 9 with the fine-tuned model and 9 with the model additionally trained using CPO. In each case, 3 different temperatures are used. A GEC model and post-processing rules, as described in Section 3.5, are applied to all translations before COMET-MBR selects the top translation candidate. Finally, post-processing rules are applied again to the translated paragraph before document translations are constructed from the paragraphs.

5 Results

In the WMT25 general translation task, automatic evaluation of participating systems was carried out using three families of evaluation methods: LLM-as-a-Judge (reference-less), Trained reference-based metrics and Trained Quality Estimation (QE).

The results, reported in Kocmi et al. (2025b), are given in Table 3. CometKiwi-XL (Rei et al., 2023) belongs to the *Trained Quality Estimation* family of evaluation methods, GEMBA-ESA (Kocmi and Federmann, 2023) to the *LLM-as-a-Judge* family and MetricX (Juraska et al., 2024) and XCOMET-XL (Guerreiro et al., 2024) are *Trained reference-based metrics*.

While 12 systems out of 33 score higher on average than our system using automatic metrics, and 9 systems score higher than us in the human evaluation, we have the second smallest model in terms of parameters and a smaller model than all higher scoring ones, at least those where the size is known. We score above our average on the two reference-based metrics, but lower when LLM-as-a-judge is used. Looking at the human evaluation results, we see that GPT 4.1 has the same order of systems for the top 5, but when the outputs are not as good, it starts to differ from the human evaluation.

6 Conclusions and Future Work

We experiment with fine-tuning very lightweight LLMs for translation and find that while our 3B parameter model can produce quite intelligible translations from English to Icelandic, they are still of considerably less quality than popular online systems and some larger language models. While we do not achieve building a model that is competitive with the best models, it is fast and can easily run locally on a modern laptop. Inference is thus inexpensive and can be fast.

We fine-tuned our model on 101k parallel sentence pairs. While we experimented with the combination of available datasets, we did not inspect why some worked better than others, e.g. why the quality was going down for the baseline data when training with more than 20k sentence pairs, but if synthetic data were added, the quality improved? What factors are at play here? We are interested in investigating that, starting with looking at diversity in the fine-tuning data.

CPO did not improve our model. We intend to look into why that was, whether it may be related to the size of the dataset or if adding more languages would be beneficial.

System Name	Params. (B)	AutoRank ↓	CometKiwi-XL ↑	GEMBA-ESA-CMDA ↑	GEMBA-ESA-GPT4.1 ↑	MetricX-24-Hybrid-XL ↑	XCOMET-XL ↑
Shy-hunyuan-MT	7	1.0	0.663	71.6	83.9	-7.5	0.543
Gemini-2.5-Pro	**?**	1.8	0.647	69.2	87.6	-7.7	0.512
GPT-4.1	**?**	1.9	0.653	70.2	84.5	-8.3	0.516
Erlendur	**?**	2.2	0.646	69.5	85.1	-8.2	0.506
TowerPlus-9B[M]	9	3.9	0.64	67.1	76.3	-8.8	0.471
ONLINE-B	**?**	4.4	0.636	66.1	73.5	-8.8	0.464
Claude-4	**?**	5.2	0.628	67.5	73.8	-10.6	0.43
TowerPlus-72B[M]	72	5.7	0.621	66.7	67.7	-10.1	0.435
TranssionTranslate	**?**	5.8	0.625	63.2	68.9	-9.1	0.43
UvA-MT	12	6.8	0.627	68.1	59.1	-11.6	0.402
CommandA-WMT	111	6.8	0.619	68.0	57.4	-11.1	0.404
GemTrans	27	7.0	0.609	65.0	59.1	-9.7	0.401
AMI	**3**	**7.4**	**0.627**	**59.6**	**58.1**	**-9.7**	**0.426**
SalamandraTA	8	8.6	0.605	61.6	53.9	-11.0	0.386
Llama-4-Maverick	400	8.8	0.587	64.7	58.8	-12.3	0.357
Mistral-Medium	**?**	9.7	0.583	65.3	51.5	-13.0	0.337
Gemma-3-27B	27	9.7	0.572	62.2	54.9	-12.4	0.364
DeepSeek-V3	671	10.5	0.547	58.0	56.6	-12.1	0.378
IRB-MT	12	11.9	0.542	61.2	47.2	-13.6	0.306
IR-MultiagentMT	**?**	12.1	0.53	60.0	51.3	-13.7	0.31
Qwen3-235B	235	13.5	0.525	60.5	41.5	-15.0	0.275
Gemma-3-12B	12	13.8	0.517	60.3	42.1	-15.4	0.268
NLLB	1	15.2	0.477	53.0	48.2	-15.0	0.27
ONLINE-G	**?**	15.8	0.477	53.4	49.2	-16.1	0.243
CommandA	111	16.2	0.475	59.0	37.4	-17.0	0.221
Llama-3.1-8B	8	24.8	0.323	42.7	24.6	-21.3	0.133
EuroLLM-9B[M]	9	25.5	0.303	32.9	9.2	-17.4	0.237
AyaExpanse-32B	32	28.0	0.275	35.2	18.4	-23.3	0.145
CommandR7B	7	30.3	0.2	23.4	9.1	-20.9	0.216
EuroLLM-22B-pre.[M]	22	30.8	0.206	26.5	13.7	-23.7	0.171
Mistral-7B	7	31.8	0.177	25.2	14.3	-24.3	0.17
Qwen2.5-7B	7	31.8	0.186	24.1	13.1	-24.3	0.174
AyaExpanse-8B	8	33.0	0.153	21.7	11.3	-24.6	0.177

Table 3: Automatic evaluation in the WMT25 General MT shared task for English→Icelandic. The table is adapted from Kocmi et al. (2025b). Our system is in bold.

Rank	**System**	**Human**
1–1	Human	87.5
2–2	Gemini-2.5-Pro	77.6
3–4	Erlendur	68.3
3–4	GPT-4.1	68.0
5–5	Shy-hunyuan-MT	63.2
6–6	TowerPlus-9B[M]	57.4
7–7	ONLINE-B	51.8
8–10	Claude-4	47.8
8–10	TowerPlus-72B[M]	46.3
8–10	TranssionTranslate	46.2
11–11	**AMI**	**39.9**
12–12	GemTrans	34.8
13–14	SalamandraTA	31.3
13–15	UvA-MT	30.6
14–15	CommandA-WMT	29.0
16–16	NLLB	24.1
17–17	IRB-MT	20.7
18–18	Gemma-3-12B	16.5
19–19	Llama-3.1-8B	10.5

Table 4: Human evaluation in the WMT25 General MT shared task for English→Icelandic. The table is adapted from Kocmi et al. (2025a). Our system is in bold.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere,

Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Starkaður Barkarson and Steinþór Steingrímsson. 2024. Icelandic Gigaword Corpus (IGC-2024ext) - unannotated version. CLARIN-IS.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022a. Evolving large text corpora: Four versions of the Icelandic Gigaword corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2371–2381, Marseille, France. European Language Resources Association.

Starkaður Barkarson, Steingrímsson Steinþór, Þórdís Dröfn Andrésdóttir, Hildur Hafsteinsdóttir, Finnur Ágúst Ingimundarson, and Árni Davíð Magnússon. 2022b. Icelandic Gigaword Corpus (IGC-2022) - unannotated version. CLARIN-IS.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Atli Jasonarson, Hinrik Hafsteinsson, Bjarki Ármannsson, and Steinþór Steingrímsson. 2024. Cogs in a Machine, Doing What They're Meant to Do – the AMI Submission to the WMT24 General Translation Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 253–262, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q. Jiang et al. 2024. Mixtral of experts.

Hildur Jónsdóttir and Anton Karl Ingason. 2020. Creating a Parallel Icelandic Dependency Treebank from Raw Text to Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2924–2931, Marseille, France. European Language Resources Association.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103. Springer.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet.

In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775. Association for Computational Linguistics.

Julia Kreutzer et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2025. Efficient Continual Pretraining of LLMs for Low-resource Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 304–317, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuyo Nakatani. 2010. Language detection library for java.

OpenAI et al. 2024. GPT-4 Technical Report.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson and Starkaður Barkarson. 2021. ParIce: English-Icelandic parallel corpus (21.10). CLARIN-IS.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, LREC 2018, pages 4361–4366, Miyazaki, Japan.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. Filtering Matters: Experiments in Filtering Training Sets for Machine Translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Steinþór Steingrímsson, Einar Freyr Sigurðsson, and Björn Halldórsson. 2024. Evaluating Capabilities of MT Systems in Translating Idiomatic Expressions Using a Specialized Dataset. In *CLARIN Annual Conference Proceedings 2024*, Barcelona, Spain.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and Scalable Sentence Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Hugo Touvron et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & Play Modules and Adaptive Rejection for Quality Translation at Scale. In *The Thirteenth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Forty-first International Conference on Machine Learning*.

Ahmet Üstün et al. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model.

A Hyperparameters

We used Accelerate and DeepSpeed for continued pre-training and fine-tuning.

A.1 Continued Pre-Training

Listing 1: Training hyperparameters

```
  max_steps: 150000
  learning_rate: 2e-5
  weight_decay: 0.01
  gradient_accumulation_steps: 4
  lr_scheduler_type: cosine
  warmup_ratio: 0.01
  per_device_train_batch_size: 4
  per_device_eval_batch_size: 4
  fp16: true
  seed: 42
  max_new_tokens: 256
  max_source_length: 256
  save_strategy: steps
  save_steps: 15000
```

Listing 2: DeepSpeed configuration for CPT

```
deepspeed_config:
  gradient_accumulation_steps: 4
  gradient_clipping: 1.0
  zero_stage: 2
  mixed_precision: fp16
distributed_type: DEEPSPEED
num_processes: 8
num_machines: 1
```

A.2 Fine-tuning

Listing 3: Fine-tuning hyperparameters

```
  num_train_epochs: 1
  learning_rate: 2e-5
  weight_decay: 0.01
  gradient_accumulation_steps: 4
  lr_scheduler_type: inverse_sqrt
  warmup_ratio: 0.01
  per_device_train_batch_size: 4
  per_device_eval_batch_size: 4
  fp16: true
  seed: 42
  max_new_tokens: 256
  max_source_length: 256
  num_beams: 5
```

Listing 4: DeepSpeed configuration for fine-tuning

```
deepspeed_config:
  gradient_accumulation_steps: 4
  gradient_clipping: 1.0
  zero_stage: 2
  mixed_precision: fp16
distributed_type: DEEPSPEED
num_processes: 2
num_machines: 1
```

KIKIS at WMT 2025 General Translation Task

Koichi Iwakawa[1], Keito Kudo[1,2], Subaru Kimura[1], Takumi Ito[1], Jun Suzuki[1,2]
[1]Tohoku University, [2]RIKEN

Abstract

We participated in the constrained English–Japanese track of the WMT 2025 General Machine Translation Task. Our system collected the outputs produced by multiple subsystems, each of which consisted of LLM-based translation and reranking models configured differently (e.g., prompting strategies and context sizes), and reranked those outputs. Each subsystem generated multiple segment-level candidates and iteratively selected the most probable one to construct the document translation. We then reranked the document-level outputs from all subsystems to obtain the final translation. For reranking, we adopted a text-based LLM reranking approach with a reasoning model to take long contexts into account. Additionally, we built a bilingual dictionary on the fly from the parallel corpus to make the system more robust to rare words.

1 Introduction

This paper describes KIKIS's submission to the WMT 2025 General Machine Translation Shared Task (Kocmi et al., 2025a,b). We participated in the constrained track for the English–Japanese (En→Ja) direction. Given limited computational resources and the rapid pace of open-source LLM releases, we aimed to build a system that produced high-quality translations without additional training. In particular, we aimed to detect and correct residual errors, such as mismatched numbers or dates, missing key terms, and unnatural phrasing, in otherwise strong translations produced by LLM-based MT systems. To this end, we adopted a multi-stage LLM-based reranking pipeline that selected the best translation from candidate outputs. This paper provides a detailed description of our submitted system. We also report post-evaluation results that demonstrate the effectiveness of each component of our system.

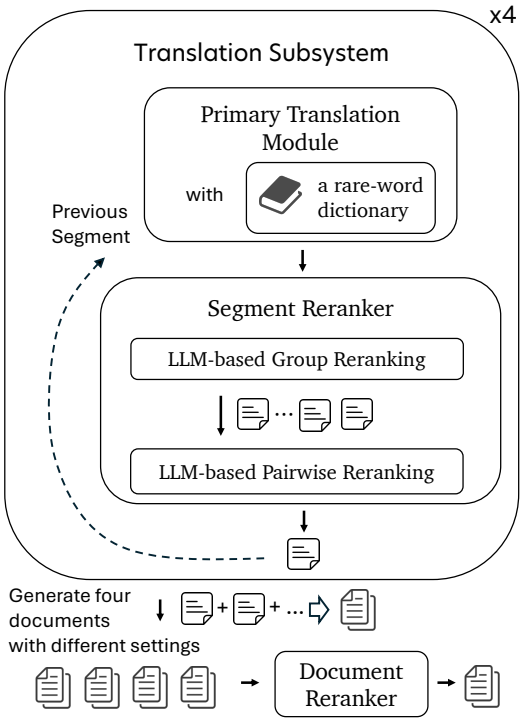


Figure 1: Overview of our final submission system.

2 System overview

Figure 1 provides an overview of our system. Our system consisted of four translation subsystems and a document reranker. We first aggregated outputs from the four subsystems, which differed in configuration (e.g., prompting strategies and context sizes). We then performed document-level reranking to select the best translation from the combined candidates. Within each subsystem, the primary translation module generated multiple segment-level hypotheses. The segment reranker iteratively filtered the candidate set via a tournament-style process to select the most plausible hypothesis. Algorithm 1 shows the pseudocode for the subsystem. Below, we describe three components: the primary translation module (§ 3), the segment reranker (§ 4), and the document reranker (§ 5).

Proceedings of the Tenth Conference on Machine Translation, pages 705–722
November 8-9, 2025

Algorithm 1 Pseudocode for the subsystem.

Require: $s_{1:T}$: T source segments
Require: τ: switch to Pairwise reranking when the number of candidates is $\leq \tau$
Require: g: group size for Group reranking
Require: m: context size
Require: `Partition`(H, n): sequentially partition set H into chunks of size n
function SUBSYSTEM($\{s_t\}_{t=1}^{T}$)
  for $t = 1$ **to** T **do**
    $H_t \leftarrow \mathrm{MT}(s_{t-m:t},\ h^{+}_{t-m:t-1})$ ▷ § 3
    $h^{+}_{t} \leftarrow$ RERANK($H_t,\ s_{t-m:t},\ h^{+}_{t-m:t-1}$) ▷ § 4
  end for
  return $h^{+}_{1:T}$
end function

function RERANK($H_t,\ s_{t-m:t},\ h^{+}_{t-m:t-1}$)
  while $|H_t| > 1$ **do**
    if $|H_t| > \tau$ **then**
      ▷ Group reranking mode (§ 4.1)
      $b \leftarrow g$
      $f_{\text{rerank}} \leftarrow$ GROUPRERANK
    else
      ▷ Pairwise reranking mode (§ 4.2)
      $b \leftarrow 2$
      $f_{\text{rerank}} \leftarrow$ PAIRWISERERANK
    end if
    $H_{\text{next}} \leftarrow [\,]$
    for B **in** PARTITION($H_t,\ b$) **do**
      $\hat{h}_t \leftarrow f_{\text{rerank}}(B,\ s_{t-m:t},\ h^{+}_{t-m:t-1})$
      $H_{\text{next}} \leftarrow H_{\text{next}} \,\|\, [\hat{h}_t]$
    end for
    $H_t \leftarrow H_{\text{next}}$
  end while
  return $H_t[1]$
end function

3 Primary translation module

We used plamo-2-translate (Imajo et al., 2025) as our base model. Plamo-2-translate is an LLM-based translation system with a hybrid architecture that combined Mamba (Gu and Dao, 2024) and Transformers (Vaswani et al., 2017). With this model, we generated 32 hypotheses for each source segment. The decoding hyperparameters are listed in Table 5. To further improve accuracy and naturalness at the document level, we combined three prompting strategies: vocabulary prompting, style prompting, and context prompting. We describe each strategy below.

3.1 Vocabulary prompting

To improve robustness to rare words, we dynamically constructed a bilingual dictionary from parallel corpora and used it as a prompt for the base model.

The dictionary was built in four steps:

- **Term extraction**: Candidate terms (e.g., named entities) were extracted from source sentences using Qwen3-8B (Qwen Team, 2025).
- **Retrieval**: Sentence pairs were retrieved from the parallel corpora whose source side contained the extracted term.
- **Translation-pair extraction**: Qwen3-8B was used to identify the translation of each term in the target sentence, and the term-level pairs were recorded.
- **Cleaning**: Pairs that were likely to be incorrectly extracted were discarded.

When a source segment in the test set contained any extracted terms, the corresponding dictionary entries were included in the prompt to the base model. In total, the dictionary comprised 365 English entries with an average of 1.9 Japanese translations per entry. See appendix C for further details (e.g., list of parallel corpora and filtering criteria).

3.2 Context prompting

To maintain document-level consistency of named entities and overall style, we translated each segment with the preceding context. For the current source segment s_t, the prompt to the base model included the m source segments $s_{t-m:t-1}$ and the previously selected hypotheses $h^{+}_{t-m:t-1}$ from the segment reranker (§ 4). During decoding, we enforced $h^{+}_{t-m:t-1}$ via forced-decoding and then generated the output for s_t.

Formally, context-prompted decoding is defined as:

$$H_t = \left\{ h_t^{(k)} \sim P_\theta\left(\cdot \mid s_{t-m:t}, h^{+}_{t-m:t-1}\right) \right\}_{k=1}^{K}, \quad (1)$$

where H_t is the set of K sampled hypotheses for the current segment, $h_t^{(k)}$ is the k-th sample, and $P_\theta(\cdot \mid s_{t-m:t}, h^{+}_{t-m:t-1})$ denotes the base model's conditional distribution.

3.3 Style prompting

We controlled the translation style based on the domain of the source document. Specifically, we prompted the base model to produce either the polite (“です/ます”) or the plain (“だ/である”) style based on the document domain. We enforced the plain style for literary and news texts and left the style unspecified for social and speech texts. We further included domain-specific instructions to keep the writing appropriate for each domain. [1]

4 Segment Reranker

This module selected plausible hypotheses from the segment-level outputs of the primary translation module via a tournament-style process. We applied two reranking stages with different granularities in sequence. In the early stage, we grouped candidates into batches of at least three and performed coarse filtering within each batch (GROUPRANK in Algorithm 1). After reducing the pool, we performed pairwise comparisons among the remaining candidates to select the final hypothesis (PAIRWISECOMPARE in Algorithm 1). Inspired by (Sun et al., 2023), we adopted a text-based LLM reranking approach using the reasoning model Qwen3-8B (Qwen Team, 2025).

4.1 Group reranking

This module selected a plausible hypothesis from a set of candidate translations. Its role was to roughly filter out low-quality outputs and narrow the candidate set. Concretely, the model received the m previous source segments and the current one, denoted $s_{t-m:t}$, the contexts of confirmed hypotheses from previous iterations $h^{+}_{t-m:t-1}$, and the current subset of candidate hypotheses generated by the primary translation module $H'_t \subseteq H_t$. From these inputs, it selected a plausible hypothesis $\hat{h}_t \in Ht'$. Formally,

$$\hat{h}_t = \mathrm{LLM_g}\left(s_{t-m:t}, h^{+}_{t-m:t-1}, H'_t\right), \qquad (2)$$

where $\mathrm{LLM_g}$ was the LLM instructed to perform group reranking, which returned one hypothesis from H'_t. The prompt template is given in appendix D.

[1] Due to terms and conditions, we cannot include the exact prompt format here. For details about the prompts, please refer to https://translate-demo.plamo.preferredai.jp/contact.

4.2 Pairwise reranking

Given the hypotheses returned by group reranking, this module performed pairwise comparisons to select the most plausible translation. As with group reranking, we used an LLM for hypothesis selection; however, here the LLM compared pairs of hypotheses. To mitigate positional bias (Liu et al., 2024; Wang et al., 2024), we prompted the LLM with each pair in both orders. If the two decisions conflicted, we selected one of the two hypotheses uniformly at random.

Formally, we expressed this as:

$$\hat{h}_t = \arg\max_{h \in H'_t} \sum_{(u,v)\in \mathrm{Perm}_2(H'_t)} \mathbf{1}\left\{\mathrm{LLM_p}\left(s_{t-m:t}, h^{+}_{t-m:t-1}, u, v\right) = h\right\}, \qquad (3)$$

Here, $\mathrm{Perm}_2(H'_t)$ denoted the set of all ordered pairs of distinct elements from H'_t. $\mathrm{LLM_p}$ was the LLM instructed to perform pairwise reranking; it returned one of the two given hypotheses, u or v. $\mathbf{1}$ denoted the indicator function. The prompt template is provided in appendix D.

5 Document reranker

We reranked document-level translation candidates generated by N_{sub} subsystems using Qwen3-8B. We performed pairwise comparisons over all ordered pairs of candidates and selected the most plausible translation. As in the segment reranking, each pair was evaluated in both input orders to deal with positional bias.

We let the set of document-level hypotheses be $\mathcal{D}_{\mathrm{hyp}} = \{D^{(1)}_{\mathrm{hyp}}, \ldots, D^{(N_{\mathrm{sub}})}_{\mathrm{hyp}}\}$, where each $D^{(i)}_{\mathrm{hyp}}$ was a complete translated document produced by a subsystem. Let the source document be $D_{\mathrm{src}} = s_{:}$ (where : denoted all source segments). We selected the final document $\hat{D}_{\mathrm{hyp}}$ by counting pairwise wins:

$$\hat{D}_{\mathrm{hyp}} = \arg\max_{d \in \mathcal{D}_{\mathrm{hyp}}} \sum_{(u,v)\in \mathrm{Perm}_2(\mathcal{D}_{\mathrm{hyp}})} \mathbf{1}\{\mathrm{LLM_d}(D_{\mathrm{src}},\, u,\, v) = d\}, \qquad (4)$$

where $\mathrm{Perm}_2(\mathcal{D}_{\mathrm{hyp}})$ denoted all ordered pairs (u, v) with $u \neq v$, and $\mathrm{LLM_d}$ was the model for document-level reranking that returned one of the two inputs, u or v. The prompt template is given in appendix D.

	Vocab prompt	Context size	Group size	Document reranking	MetricX↓ w/ ref	MetricX↓ w/o ref	XCOMET↑ w/ ref	XCOMET↑ w/o ref
Primary translation module only								
(a)		0			5.62	6.02	0.522	0.502
(b)	✓	0			5.60	6.03	0.520	0.500
(c)		2			5.59	6.01	0.528	0.498
Subsystems								
(d)		2	4		5.50	5.90	0.547	0.519
(e)	✓	2	4		5.40	5.82	0.552	0.521
(f)	✓	4	4		5.58	5.95	0.542	0.508
(g)	✓	2	8		5.49	5.86	0.548	0.520
Final submission system								
(h)	✓	2-4	4-8	✓	5.51	5.92	0.551	0.517

Table 1: Post evaluation results.

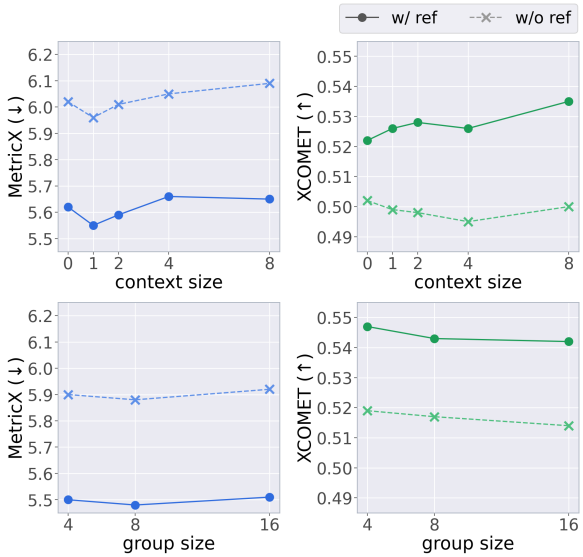


Figure 2: Automatic evaluation scores with varying context sizes (top row) and group sizes (bottom row). The left column shows MetricX results (lower is better) and the right column shows XCOMET results (higher is better). Solid lines indicate reference-based (w/ ref) evaluation, and dashed lines indicate reference-free (w/o ref) evaluation.

6 Post-evaluation

We conducted a post-evaluation to assess the contribution of each component in our system.

6.1 Experimental setup

We reported the performance of our final submission. The final system consisted of four subsystems, and we also reported the performance of each subsystem (before applying document-level reranking). As ablation studies, we evaluated the effects of removing the segment reranker, enabling or disabling vocabulary prompting, and varying the context size (m in § 3.2) and the group size (g in Algorithm 1). In the no-reranking setting (using only the primary translation module), we generated 32 translation candidates for each source segment (as in Section 3) and selected the highest-probability candidate.

We used XCOMET-XL (Guerreiro et al., 2024)[2] and MetricX-24 (XL) (Juraska et al., 2024)[3] as automatic evaluation metrics. We evaluated our approach in both reference-based and reference-free (quality estimation) settings.

6.2 Results and discussion

Table 1 shows the post-evaluation results.

Effect of reranking. Comparing configurations (c) and (d) in Table 1, we observed that segment-level reranking (§ 4) improved performance, suggesting that the LLM selects better translations. By contrast, in our experimental setting, document-level reranking (h)(§ 5) did not surpass configuration (d), which was the best-performing subsystem before reranking. This may be partly due to the much longer input at the document level, which could make the reranking task more challenging for the LLM.

Effect of vocabulary prompting. In Table 1, the comparisons between (a) and (b) and between (d) and (e) showed no consistent effect of vocabulary prompting on evaluation metrics. Qualitatively, however, as shown in Table 2, vocabulary prompting improved translations of domain-specific terms; for example, "facial scrub" is translated correctly.

Effect of context sizes. Figure 2 shows the relationship between context size and performance. According to the automatic evaluation metrics, we

[2] https://huggingface.co/Unbabel/XCOMET-XL
[3] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

Source	you're gonna cleanse get rid of all the grease and makeup from your face and then you're gonna use your **facial scrub**
Reference	肌の上に。さて、こちらがその**フェイシャルスクラブ**です。このスクラブを使う際に、まず最初に忘れずに顔をクレンジングします。それが第一のステップです。
Translation	
(d) w/o vocab	ここで重要なのは、フェイススクラブを使用する前には必ず洗顔を行うことです。
(e) w/ vocab	まず顔の皮脂やメイクを完全に落とします。その後、このフェイシャルスクラブを使います。

Table 2: Qualitative output evaluation: vocabulary usage ((d) vs. (e)). Incorporating a predefined vocabulary list (e) ensured that the translation matched the reference term "フェイシャルスクラブ", in contrast to the variant "フェイススクラブ" produced without the vocabulary (d).

Source	Segment t:	First job is taking out the floor ⋯ cover that in linoleum. The job is shoddy ⋯ I dunno how to do all that. The roof will remain cold and metal ⋯
	Segment $t+1$:	This chair works very well in the van ⋯ should ideally be fastened better ⋯ flat as a bed.
Reference	Segment t:	「床を外して下の収納スペースにアクセスし、そこにリノリウムを貼ること。…正直言って雑な仕事。…そんな技術はまったくない。…屋根は金属むき出しで冷たいまま。」
	Segment $t+1$:	「床にちゃんと固定できてはいないけど… フラットなベッドになる。」
(a) Context size = 0	Segment t:	「床板を取り外し…敷くことだ。…この作業は粗雑な仕上がりになってしまう。…方法がわからない。…屋根は金属のまま放置することになる。」
	Segment $t+1$:	「椅子は理想的には固定すべきですが…ベッドとしても使える状態になります。」
(c) Context size = 2	Segment t:	「床を撤去し…敷くことです。…この作業は雑な仕上がりです。…方法が分かりません。…屋根は金属のままとなります。」
	Segment $t+1$:	「椅子は固定方法を改善すべきですが…ベッドとしても使用可能です。」

Table 3: Qualitative evaluation of outputs: effect of context size ((a) vs. (c)). (a) Without previously translated hypotheses as context, translations mix polite ("です/ます") and plain ("だ/である") sentence endings across adjacent segments. (c) With previously translated hypotheses provided as context, the sentence-ending style remains consistent, avoiding style shifts between segments.

did not observe a consistent trend in performance as the context size changed. Qualitatively, however, as shown in Table 3, adding previously translated hypotheses as in-document context helped keep a consistent style across the document, either polite ("です/ます") or plain ("だ/である").

Effect of group size. Figure 2 shows how group size affected segment-level reranking. Intuitively, larger groups made it harder to select the best candidate. At the same time, they allowed us to prune more candidates per group, which sped up the system. Across the tested group sizes, XCOMET and MetricX scores dropped by less than 0.05 points. Therefore, there was room to either speed up the final submission system or use the saved time to generate more translation candidates from the primary translation module within the same time budget.

7 Conclusion

We described the KIKIS submission to the WMT 2025 General Machine Translation Shared Task. We participated in the constrained track for the English–Japanese (En→Ja) direction. Our system consisted of four translation subsystems and a document reranker. Each subsystem combined an MT model with an LLM-based segment reranker. We aggregated the outputs from the four subsystems and then applied document-level reranking to select the final translation.

Acknowledgments

We would like to thank the members of the Tohoku NLP Group for their cooperation and feedback throughout the course of this research. This work was supported by Moonshot R&D Grant Number JPMJMS2011-35 (fundamental research), JST BOOST Japan Grant Number JPMJBS2421 and

JSPS KAKENHI Grant Number JP25KJ0615.

Contributions

Koichi Iwakawa conducted hyperparameter search for LLM-based translation models and performed post-evaluations.
Keito Kudo built the foundation of the LLM-based translation system, developed the reranking system, and constructed the bilingual dictionary.
Subaru Kimura was responsible for decoding the test set samples using the submission system. He also investigated translation quality under different prompting strategies.
Takumi Ito contributed to strategic discussions for the WMT submission and provided feedback and advice on translation outputs.
Jun Suzuki supervised and coordinated the entire project.

References

Alfred V. Aho and Margaret J. Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18(6):333–340.

Inc. Baobab. 2024. baobab_coco_evaluate_caption_24. Initial commit on 2024-11-25; updates on 2024-12-19 and 2024-12-21.

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314, Toronto, Canada. Association for Computational Linguistics.

Bianka Buschbeck and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation. In *Proceedings of the 7th Workshop on Asian Translation*, pages 160–169, Suzhou, China. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. Ccaligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vincent J. Felitti, Robert F. Anda, Dale Nordenberg, David F. Williamson, Alison M. Spitz, Valerie Edwards, Mary P. Koss, and James S. Marks. 1998. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The adverse childhood experiences (ace) study. *American Journal of Preventive Medicine*, 14(4):245–258. The original Adverse Childhood Experiences Study.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

GENIAC Team Ozaki. 2024. Wikihownfqa-ja_cleaned. Created on 2024-05-10; license CC BY 4.0.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kazuma Hashimoto, Raffaella Buschiazzo, James Bradbury, Teresa Marshall, Richard Socher, and Caiming Xiong. 2019. A high-quality multilingual dataset for structured documentation translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 116–127, Florence, Italy. Association for Computational Linguistics.

Yuta Hayashibe. 2023. megagonlabs/instruction_ja. https://github.com/megagonlabs/instruction_ja. GitHub repository.

Kentaro Imajo, Masanori Hiano, Kento Nozawa, and Chubachi Kaizabro. 2025. Plamo translate: Development of a large language model specialized for translation (original japanese title: "PLaMo Translate: 翻訳特化大規模言語モデルの開発"). Technical report, Preferred Networks, Inc.

Tatsuya Ishisaka, Masao Utiyama, Eiichiro Sumita, and Kazuhide Yamamoto. 2009. Building a large scale japanese-english open source parallel corpus. *IPSJ SIG Technical Report*, 2009(1):1–6. Also appears in SLP Vol. 2009-SLP-76.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Katsuhiko Toyama. 2009. Japanese law translation project. https://www.kl.i.is.nagoya-u.ac.jp/told/index.html. Accessed: 2025-08-10.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025b. Preliminary ranking of wmt25 general machine translation systems.

Keito Kudo, Hiroyuki Deguchi, Makoto Morishita, Ryo Fujii, Takumi Ito, Shintaro Ozaki, Koki Natsumi, Kai Sato, Kazuki Yano, Ryosuke Takahashi, Subaru Kimura, Tomomasa Hara, Yusuke Sakai, and Jun Suzuki. 2024. Document-level translation with LLM reranking: Team-J at WMT 2024 general translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 210–226, Miami, Florida, USA. Association for Computational Linguistics.

Kurohashi-Kawahara Lab. and NICT. 2011. JEC Basic Sentence Data.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, Ryoko Tokuhisa, Ana Brassard, and Kentaro Inui. 2022. Chat translation error detection for assisting cross-lingual communications. In *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems*, pages 88–95, Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Andrew Merritt, Chenhui Chu, and Yuki Arase. 2020. A corpus for english-japanese multimodal neural machine translation with comparable sentences. *CoRR*, abs/2010.08725.

Mitsua. 2024. Wikidata parallel descriptions (en–ja). Initial commit on 2024-05-13; updated 2024-05-17. Generated from Wikidata dump 2024-05-06; 1,570,685 rows.

Rei Miyata. 2024. Mtpedocs. https://github.com/tntc-project/MTPEdocs. Tntc-project GitHub repository.

Mozilla Contributors. 2005–2025. Mdn web docs. Online. Web platform documentation and learning resource.

Atsushi Nakajima. 2022. fungi_indexed_mycological_papers_japanese. Compiled summaries, tags, reported and compared species from the Daikinrin website.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: Asian

scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2204–2208, Portorož, Slovenia. European Language Resources Association (ELRA).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. Tufs asian language parallel corpus (talpco). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*, pages 436–439, Okayama, Japan. Association for Natural Language Processing. Japanese-based parallel corpus for Burmese, Malay, Indonesian, and English.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Matīss Rikters, Ryokan Ri, Tong Li, and Toshiaki Nakazawa. 2019. Designing the business conversation corpus. In *Proceedings of the 6th Workshop on Asian Translation*, pages 54–61, Hong Kong, China. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Haiyue Song, Raj Dabre, Atsushi Fujita, and Sadao Kurohashi. 2020. Coursera corpus mining and multistage fine-tuning for improving lectures translation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3640–3649, Marseille, France. European Language Resources Association.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 685–691.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Yasuhito Tanaka, Jim Breen, and Paul Blay. 2001. Tanaka corpus. Originally compiled at Hyogo University, now part of Tatoeba Project. Japanese-English parallel sentence corpus, maintained by Tatoeba Project.

Tatoeba Community. 2006–2025. Tatoeba: Collection of sentences and translations. Online collaborative platform. Community-driven parallel corpus with over 12.6M sentences in 426 languages.

Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA). EUbookshop is part of the OPUS collection.

TLDR Pages Community. 2013–2025. Tldr pages: Collaborative cheatsheets for console commands. GitHub repository. Community-maintained help pages for command-line tools with 56,000+ GitHub stars.

Hayato Tsukagoshi. 2024. hpprc/honyaku. Repository owner: hpprc; dataset card indicates CC BY-SA 4.0.

Masao Utiyama. 2019. Paranatcom — parallel english-japanese abstract corpus made from nature communications articles.

Masao Utiyama and Mayumi Takahashi. 2023. English-japanese translation alignment data. Page last updated on 2016-08-25; dataset originally released in 2003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. 2024. Eliminating position bias of language models: A mechanistic approach. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

	Parameters
plamo-2-translate	9.5 B
Qwen3-8B	8.2 B
Total	17.7 B

Table 4: Model parameter sizes.

plamo-2-translate	
temperature	0.8
top-p	0.9
Qwen3-8B	
temperature	0.6
top-p	0.95
top-k	20
min-p	0.0
Output format	structured outputs (json)

Table 5: Decoding hyperparameters.

A Decoding hyperparameters

Table 5 lists the decoding hyperparameters used in our system. For Qwen3-8B decoding, we adopted the officially recommended settings[4]. We used vllm (Kwon et al., 2023) to decode translation candidates and to run the reranking steps.

B Parameter count

Table 4 shows the parameter counts of the models used in our translation system. Our system satisfied the constrained track limit of at most 20B total parameters. We did not fine-tune any model and used the publicly available model parameters as provided.

C Detail of bilingual dictionary construction

This section described the detailed procedure for constructing the bilingual dictionary described in § 3.1.

Term extraction. We first extracted named entities and technical terms from the source sentences in the test set. We used Qwen3-8B (Qwen Team, 2025) to perform term extraction. The prompt we gave to Qwen3-8B for term extraction is shown in List D. After extraction, we filtered out terms that met any of the following conditions:

- The term fell within the top 66% by frequency within the test set.
- The term was tokenized as a single token by the plamo-2-translate tokenizer.
- The term was included in the top 100,000 words of the English word frequency list from `wordfreq` (Speer, 2022).

We applied this filtering because terms that satisfied these conditions were frequent and were expected to be translated correctly by the base model without explicit vocabulary prompts.

Retrieval. We retrieved all parallel sentence pairs from the corpora whose source side contained any of the extracted terms. We used the Aho–Corasick algorithm (Aho and Corasick, 1975) for efficient multi-pattern matching. Table 6 lists the parallel corpora used as sources for the bilingual dictionary. For several corpora, we computed LaBSE (Feng et al., 2022) sentence embeddings and filtered out sentence pairs whose semantic similarity was outside the range [0.7, 0.96]. Pairs with LaBSE similarity below 0.7 were removed due to likely low semantic alignment between source and target. Pairs with LaBSE similarity above 0.96 were removed because they were likely to be nearly identical (copying) or noisy. The Team-J (WMT2024) bitext dataset and development dataset, which we used in last year' s submission and which was built from open data sources, was included and was filtered with the same LaBSE-based criterion. See Kudo et al. (2024) for more details.

Translation-pair extraction. We then used Qwen3-8B to extract candidate target terms from the retrieved parallel sentences. The prompt given to the model was shown in List D. This process yielded multiple candidate target-term patterns for each source term.

Cleaning. Some extracted target terms were noisy or incorrect due to model errors. Therefore, for each source term, we retained only the most frequently extracted target terms. Concretely, we ranked the extracted target terms by occurrence frequency and kept the top 30This frequency-based filtering reduced noise and favored stable translations that appeared repeatedly in the parallel data.

[4] https://huggingface.co/Qwen/Qwen3-8B#best-practices

	Filtering	Size
Team-J WMT2024 bitext dataset (Kudo et al., 2024)	✓	22,899,294
Team-J WMT2024 development dataset (Kudo et al., 2024)	✓	18,113
BSD (Rikters et al., 2019)		808
BPersona-chat (Li et al., 2022; Sugiyama et al., 2023)		2,940
MTPEdocs (Miyata, 2024)		1045
CourseraParallelCorpusMining (Song et al., 2020)		53,166
Flickr30kEnt-JP		155,070
JSICK (Yanaka and Mineshima, 2022)		18,854
IWSLT2017 (Cettolo et al., 2017)		9,340
Software Documentation Data Set for Machine Translation (Buschbeck and Exel, 2020)	✓	7,745
localization-xml-mt (Hashimoto et al., 2019)	✓	82,546
Asian Language Treebank Parallel Corpus (Thu et al., 2016)		20,101
ParaNatCom (Utiyama, 2019)		507
JEC Basic Sentence Data (Kurohashi-Kawahara Lab. and NICT, 2011)		4,769
Japanese Law Translation (Katsuhiko Toyama, 2009)		75,930
English–Japanese Translation Alignment Data (Utiyama and Takahashi, 2023)		42,738
Tanaka Corpus (Tanaka et al., 2001)		147,865
honyaku (Tsukagoshi, 2024)		33
TALPCo (Nomoto et al., 2018)		1,372
WikiHowNFQA-ja-en (Bolotova-Baranova et al., 2023; GENIAC Team Ozaki, 2024)		9,584
fungi_indexed_mycological_papers_japanese (Nakajima, 2022)		12,744
baobab_coco_evaluate_caption_24 (Baobab, 2024)		50
ACES (Felitti et al., 1998)		430
MASSIVE (FitzGerald et al., 2023)		11,514
ea-mt-benchmark (Conia et al., 2024)		5,108
JaEnCOCO (Merritt et al., 2020)		461
instruction_ja (Hayashibe, 2023)		669
ASPEC (Nakazawa et al., 2016)	✓	1,465,019
Large Scale Japanese-English Open Source Parallel Corpus (Ishisaka et al., 2009)	✓	180,709
CCAligned (El-Kishky et al., 2020)	✓	11,544,406
CCMatrix (Schwenk et al., 2021)	✓	28,827,886
EUbookshop (Tiedemann, 2012)	✓	81
GNOME (Tiedemann, 2012)	✓	37
HPLT (de Gibert et al., 2024)	✓	14,759,898
KDE4 (Tiedemann, 2012)	✓	81,316
MDN Web Docs (Mozilla Contributors, 2005–2025)	✓	65,621
NLLB (NLLB Team et al., 2022)	✓	28,827,883
PHP (Tiedemann, 2012)	✓	3,109
QED (Tiedemann, 2012)	✓	24,289
Tanzil (Tiedemann, 2012)	✓	53,202
Tatoeba (Tatoeba Community, 2006–2025)	✓	189,386
XLEnt (El-Kishky et al., 2021)	✓	2,577,352
tldr-pages (TLDR Pages Community, 2013–2025)	✓	720
Wikidata parallel descriptions (Mitsua, 2024)	✓	860,742
Total		113,044,452

Table 6: A list of the parallel corpus used for bilingual dictionary construction. Entries with a ✓ in the filtering column indicate that LaBSE-based data filtering was applied; the Size column shows the number of sentence pairs.

D Prompt lists

Term extraction. The prompt template used for term extraction described in § 3.1 was as follows. The source segment from the test set was embedded in <|INPUT_TEXT|>.

```
# Named Entity Recognition and
Technical Term Extraction Instructions

Extract named entities and technical
terms from the following text.

## Processing Steps

1. **Named Entity Extraction**
   Extract expressions that denote
   specific entities belonging to the
   following categories:

   - **PERSON**: Individual names,
   fictional character names
   - **ORGANIZATION**: Companies,
   government agencies, organizations,
   teams, etc.
```

```
   - **LOCATION**: Countries, cities,
   regions, buildings, natural
   features, etc.
   - **PRODUCT**: Products, services,
   software, titles of works, etc.
   - **EVENT**: Conferences, festivals,
   competitions, historical events,
   etc.
   - **MISC**: Other specific entities
   not classified above

   **Important**:
   - Exclude commonly known entities
   that appear frequently in general
   texts (e.g., "United States",
   "Japan", "Google", "Microsoft",
   "China", "Tokyo", etc.)
   - Focus on extracting entities that
   are specific and distinctive to the
   given text
   - **Prioritize entities that may
   pose translation challenges**, such
   as:
     * Local or regional entities with
     cultural significance
     * Organizations with acronyms or
     abbreviations
     * Location names with specific
     cultural or historical context
     * Products with brand-specific
     terminology

   ※ Numbers and dates should not be
   included in the extraction

2. **Technical Term Extraction**
   Extract domain-specific terminology
   including:

   - **TECHNICAL**: Scientific,
   medical, engineering, IT, legal,
   financial, and other field-specific
   terminology
   - **CONCEPT**: Abstract concepts,
   theories, methodologies, principles
   specific to certain domains
   - **PROCESS**: Specialized
   procedures, techniques, or methods
   used in specific fields

   **Important**:
   - Exclude commonly used technical
   terms that are widely known (e.g.,
   "computer", "internet", "software",
   "database", "algorithm" in IT
   contexts)
   - Focus on specialized or
   domain-specific terms that provide
   unique insights
   - **Prioritize terms that are
   challenging for translation**,
   including:
     * Field-specific jargon with no
     standard translation
     * Compound terms or phrases unique
     to the domain
     * Terms requiring deep domain
     knowledge for accurate translation
     * Newly coined terms or emerging
     concepts
     * Terms with specific meanings
     that differ from general usage

3. **Extraction Priority**
   Apply the following filtering
   process for both named entities and
   technical terms:

   1. First, identify all potential
   entities and terms
   2. Filter out generally
   common/well-known items
   3. Keep only distinctive and
   informative items
   4. **Translation Difficulty
   Priority**: Prioritize entities and
   terms that are likely to be
   challenging for translation:
      - Culture-specific concepts with
      no direct equivalent in other
      languages
      - Domain-specific terminology
      requiring specialized knowledge
      - Acronyms, abbreviations, and
      neologisms
      - Context-dependent expressions
      - Terms with multiple meanings
      that require disambiguation
      - Compound technical terms unique
      to specific fields

## Output Format
```

````
Please provide the output in the
following JSON format:

```json
{
  "named_entities": {
    "person": ["List of extracted
    person names"],
    "organization": ["List of extracted
    organization names"],
    "location": ["List of extracted
    location names"],
    "product": ["List of extracted
    products/services"],
    "event": ["List of extracted
    events"],
    "misc": ["List of other named
    entities"]
  },
  "technical_terms": {
    "technical": ["List of
    field-specific terminology"],
    "concept": ["List of abstract
    concepts and theories"],
    "process": ["List of specialized
    procedures and methods"]
  }
}
```

## Example

### Input example
```
Yesterday, Dr. John Smith from Tohoku
University won first place in the
WMT2025 machine translation competition
held in Suzhou, China. His team's
neural translation system outperformed
submissions from Meta AI, DeepMind, and
Microsoft Research with a BLEU score of
45.7. The competition focused on
translation tasks covering more than 10
languages, including low-resource
language pairs. The transformer
architecture with attention mechanisms
proved crucial for handling
morphologically complex languages.
```

### Output example
```json
{
  "named_entities": {
    "person": ["John Smith"],
    "organization": ["Tohoku
    University", "Meta AI", "DeepMind",
    "Microsoft Research"],
    "location": ["Suzhou"],
    "product": [],
    "event": ["WMT2025 machine
    translation competition"],
    "misc": []
  },
  "technical_terms": {
    "technical": ["neural translation
    system", "BLEU score",
    "transformer architecture",
    "attention mechanisms"],
    "concept": ["low-resource language
    pairs", "morphologically complex
    languages"],
    "process": []
  }
}
```
※ Note: "China" and common terms like
"machine translation" and "translation
tasks" are excluded from the extraction
as they are generally well-known

## Notes

- Include named entities and technical
terms in the list without duplication
- If no items are found for a specific
category, output an empty list for that
category
- Distinguish between named entities
(specific instances) and technical
terms (domain concepts)
- A term can be both a product name and
a technical term depending on context
- **Exclude entities and terms that are
commonly known or frequently used in
general discourse** (e.g., major
countries, well-known companies, basic
technical terms)
- Focus on extracting distinctive,
informative, and document-specific
entities and terms
- **From a translation perspective,
prioritize extraction of:**
````

```
  * Terms requiring cultural or
  contextual knowledge
  * Domain-specific expressions without
  established translations
  * Ambiguous terms needing
  disambiguation
  * Entities with specific
  local/regional significance

---


## Input
```
<|INPUT_TEXT|>
```
```

Translation-pair extraction. The prompt template used for translation-pair extraction described in appendix C was as follows. The source sentence selected from the parallel corpus was embedded in <<<|SOURCE_TEXT|>>>, and the target sentence selected from the parallel corpus was embedded in <<<|TARGET_TEXT|>>>. Additionally, the term contained in the source sentence that was extracted during the term extraction phase was embedded in <<<|TERM|>>>.

```
# Translation Term Extraction Task

## Task Overview
Extract the corresponding translation
for a specified term from given source
text (original) and target text
(translated) pairs.

## Input Format
The following three elements will be
provided:
- **source text**: The original text
- **target text**: The translated text
- **term**: A specific term contained
in the source text

## Processing Steps
1. Identify the specified term within
the source text
2. Analyze how that term is translated
in the target text
3. Extract the corresponding
translation **exactly as it appears**
from the target text
4. When extracting, use the character
string that actually appears in the
target text without any modifications

## Important Notes
- The source text and target text may
not necessarily have a perfect
translation relationship
- The translation corresponding to the
specified term may not exist in the
target text
- When extracting translations, do not
add speculation or corrections; use
only character strings that actually
exist in the target text
- If no corresponding translation is
found, or if it is determined that no
translation relationship exists,
output null

## Output Format
Output in the following JSON format:

```json
{
  "term": "input term",
  "term_translation": "extracted
  translation" or null
}
```

Examples

Example 1 (Normal extraction)
Input:
- source text: "The artificial
intelligence system processes data
efficiently."
- target text: "その人工知能システムは
データを効率的に処理します。"
- term: "artificial intelligence"

Output:
```json
{
  "term": "artificial intelligence",
  "term_translation": "人工知能"
}
```

Example 2 (Translation not found)
```

```
**Input:**
- source text: "The quantum computer solved the complex problem."
- target text: "その高性能コンピュータは難しい問題を解決した。"
- term: "quantum"

**Output:**
```json
{
  "term": "quantum",
  "term_translation": null
}
```

Example 3 (Unclear translation relationship)
Input:
- source text: "The new policy will be implemented next month."
- target text: "会議は来週開催されます。"
- term: "policy"

Output:
```json
{
  "term": "policy",
  "term_translation": null
}
```

Input
Source Text
<<<|SOURCE_TEXT|>>>

Target Text
<<<|TARGET_TEXT|>>>

Term
<<<|TERM|>>>
```

Pairwise reranking. The prompt template used for pairwise reranking described in § 4.2 was as follows. The surrounding source segments that provided document-level context (e.g., preceding and following segments) were embedded in <<<|SURROUNDING_CONTEXT|>>>. The translations of previously processed source segments, used as context for the current translation, were embedded in <<<|PREVIOUSLY_TRANSLATED_CONTEXT|>>>. The source sentence to be translated (or the source sentence selected from the parallel corpus) was embedded in <<<|SOURCE_TEXT|>>>. A candidate translation for the current source segment (Translation A) was embedded in <<<|TRANSLATION_A|>>>. A candidate translation for the current source segment (Translation B) was embedded in <<<|TRANSLATION_B|>>>.

```
## Task
You will be given the following information as input:
1. Source sentence (original text)
2. Surrounding context (English)
3. Already translated preceding context (Japanese)
4. Two machine translation candidates A and B (Japanese)

Evaluate both candidates and **select the better translation** based on comprehensive quality assessment.

---

### 1. Evaluation Criteria
Determine the ranking according to the following 5 criteria:
1. **Adequacy**: How accurately the meaning of the source text is conveyed
2. **Fluency**: Grammatical correctness and especially the naturalness, readability, and rhythm of the Japanese. Non-literal translations are preferred.
3. **Terminology \& Proper Nouns**: Accuracy and consistency of technical terms and proper nouns. Eliminate any inconsistency in terminology usage and avoid variation in spelling or phrasing of proper nouns.
4. **Style**: Tone, punctuation, and formatting appropriate for the purpose and audience. Choose a tone that aligns with the context—such as conversational for social media posts or literary for narrative writing.
```

````
5. **Contextual Consistency**:
Consistent expression with the source
text, its surrounding context, and the
preceding translated content

---

### 2. Output Format (JSON)
Return only a JSON object following the
schema below.
Do not include any extra keys,
comments, or trailing commas.
```json
{
  "general_comment": "<Describe the
  overall reasoning for the
  selection>",
  "comparison_results": {
    "translation_A": {
      "strengths": "<Key strengths of
      Translation A>",
      "weaknesses": "<Key weaknesses of
      Translation A>"
    },
    "translation_B": {
      "strengths": "<Key strengths of
      Translation B>",
      "weaknesses": "<Key weaknesses of
      Translation B>"
    },
    "selection_reason": "<Brief
    explanation why the selected
    translation is better>",
    "selected_translation": "<A or B>"
  }
}
```
- general_comment: Overall assessment
explaining the comparison and selection
rationale.
- comparison_results
  - translation_A/B: Analysis of each
  translation
    - strengths: Main advantages of
    this translation
    - weaknesses: Main disadvantages of
    this translation
  - selection_reason: Concise
  explanation of why the selected
  translation is superior
  - selected_translation: "A" or "B" -
  the better translation


Even if the translations are very
similar in quality, always make a
definitive selection.

---

### 3. Comparison Procedure
- In the thinking process, please
conduct a detailed comparison by:
  1. Analyzing each translation against
  all evaluation criteria
  2. Identifying specific differences
  between Translation A and Translation
  B
  3. Weighing the relative importance
  of these differences in the given
  context
  4. Making a final judgment based on
  overall quality

---

### 4. Input
Context:
```txt
<<<|SURROUNDING_CONTEXT|>>>
```
Translated Context:
```txt
<<<|PREVIOUSLY_TRANSLATED_CONTEXT|>>>
```
Source:
```txt
<<<|SOURCE_TEXT|>>>
```
Translation A:
```txt
<<<|TRANSLATION_A|>>>
```
Translation B:
```txt
<<<|TRANSLATION_B|>>>
```
````

Group reranking. The prompt template used for group reranking described in § 4.1 was as follows. The surrounding source segments that provided document-level context (e.g., preceding and following segments) were embedded in <<<|SURROUNDING_CONTEXT|>>>.

The translations of previously processed source segments, used as context for the current translation, were embedded in <<<|PREVIOUSLY_TRANSLATED_CONTEXT|>>>.
The source sentence to be translated (or the source sentence selected from the parallel corpus) was embedded in <<<|SOURCE_TEXT|>>>.

```
## Task
You will be given the following
information as input:
1. Source sentence (original text)
2. Surrounding context (English)
3. Already translated preceding context
(Japanese)
4. N machine translation candidates
(Japanese)

Evaluate each candidate and **rank them
from highest quality (rank_1) to lowest
quality (rank_N)**.

---

### 1. Evaluation Criteria
Determine the ranking according to the
following 5 criteria:
1. **Adequacy**: How accurately the
meaning of the source text is conveyed
2. **Fluency**: Grammatical
correctness and especially the
naturalness, readability, and rhythm of
the Japanese. Non-literal translations
are preferred.
3. **Terminology \& Proper Nouns**:
Accuracy and consistency of technical
terms and proper nouns. Eliminate any
inconsistency in terminology usage and
avoid variation in spelling or phrasing
of proper nouns.
4. **Style**: Tone, punctuation, and
formatting appropriate for the purpose
and audience. Choose a tone that aligns
with the context—such as
conversational for social media posts
or literary for narrative writing.
5. **Contextual Consistency**:
Consistent expression with the source
text, its surrounding context, and the
preceding translated content

---

### 2. Output Format (JSON)
Return only a JSON object following the
schema below.
Do not include any extra keys,
comments, or trailing commas.

```json
{
  "general_comment": "<Describe the
  overall reasoning for the ranking>",
  "reranking_results": {
    "rank_1": {
      "translation_id": <integer>,
      "score": <0 ~ 100>
    },
    "rank_2": {
      "translation_id": <integer>,
      "score": <0 ~ 100>
    }
    // ⋯
    ,
    "rank_N": {
      "translation_id": <integer>,
      "score": <0 ~ 100>
    }
  }
}
```

- general_comment: Overall assessment
explaining the ranking rationale.
- reranking_results
  - rank_i: rank_1 is the highest
  quality, rank_N is the lowest
  quality.
  - translation_id: The candidate
  number indicated in the input
  (Translation 1 ⇒ 1, Translation 2 ⇒
  2, ⋯).
  - score: Overall score from 0 to 100.
  Higher scores indicate better
  quality.

Even in case of ties, always determine
a definitive order and assign unique
ranks without duplicates.

```

````
### 3. Reranking Procedure
- In the thinking process, please make a final judgment by repeatedly comparing what differences exist between each pair of translation results and comparing which translation result is better.

---


### 4. Input
Context:
```txt
<<<|SURROUNDING_CONTEXT|>>>
```

Translated Context:
```txt
<<<|PREVIOUSLY_TRANSLATED_CONTEXT|>>>
```

Source:
```txt
<<<|SOURCE_TEXT|>>>
```
````

Document reranking. The prompt template used for document reranking described in § 5 was as follows. The source document to be translated was embedded in <<<|SOURCE_TEXT|>>>. A candidate translation for the current source document (Translation A) was embedded in <<<|TRANSLATION_A|>>>. A candidate translation for the current source document (Translation B) was embedded in <<<|TRANSLATION_B|>>>.

````
## High-quality translation result selection task

Given an **English source text** and two **Japanese translation candidates (A, B)**,
compare them and determine which one is superior overall, then respond in the specified format.

---

### 1. Evaluation Criteria

1. **Critical errors or unnaturalness can single-handedly determine the selection.**
2. **Adequacy** - Whether the translation accurately conveys all meaning and information from the source
3. **Fluency \& Style** - Whether it reads naturally in Japanese / Whether the style and tone match the context
4. **Terminology \& Proper Nouns** - Consistency and accuracy of technical terms and proper noun translations
5. **Consistency** - Coherence with surrounding paragraphs/sentences (tense, person, etc.)
6. **Readability** - Whether punctuation, line breaks, word order, etc. are reader-friendly

---

---

### 2. Output Format (one line per input segment)

Please output in the following JSON format:

```json
{
  "selected_translation": "A" | "B",
  "general_comment": "<1-2 sentences explaining the decisive factor>"
}
```

**Constraints**

* `selected_translation` must be either `"A"` or `"B"`.
* Do not output any characters outside the JSON (no surrounding \`\`\`, etc.).
* `general_comment` should be 2 sentences maximum.
* Always select one even if uncertain.

---

### 3. Comparison Procedure
````

````
1. Check A / B individually against the
6 evaluation criteria.
2. Prioritize significant differences
(mistranslations, fluency breakdowns,
terminology inconsistencies, etc.).
3. Specify the overall superior choice
as `selected_translation`.
4. Summarize the decisive factor
concisely.



### 4. Input

#### SOURCE:
```txt
<<<|SOURCE_TEXT|>>>
```

#### TRANSLATION A:
```txt
<<<|TRANSLATION_A|>>>
```

#### TRANSLATION B:
```txt
<<<|TRANSLATION_B|>>>
```
````

Google Translate's Research Submission to WMT2025

Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan,
Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, David Vilar
Google Translate

Abstract

Large Language Models have shown impressive multilingual capabilities, where translation is one among many tasks. Google Translate's submission to the 2025 WMT evaluation tries to research how these models behave when pushing their translation performance to the limit. Starting with the strong Gemma 3 model, we carry out supervised fine tuning on high quality, synthetically generated parallel data. Afterwards we perform an additional Reinforcement Learning step, with reward models based on translation metrics to push the translation capabilities even further. Controlling the combination of reward models, including reference-based and quality estimation metrics, we found that the behaviour of the model could be tailored towards a more literal or more creative translation style. Our two submissions correspond to those two models. We chose the more creative system as our primary submission, targetting a human preference for better sounding, more naturally flowing text, although at the risk of losing on the accuracy of the translation. It is an open question to find the sweet spot between these two dimensions, which certainly will depend on the specific domain to handle and user preferences.

1 Introduction

In this paper we present Google Translate's research submission to the General MT track for the WMT 2025 shared task. Starting with Gemma 3 (Gemma Team, 2025), a strong multilingual LLM, we focus on improving its translation capabilities through supervised fine-tuning (Section 2) and reinforcement learning (RL) (Section 3). We use a mix of human- and synthetically-generated parallel data for boosting translation performance, as well as general domain post-training data in order to mostly retain the general capabilities of the original model. Through combinations of different reward models in the reinforcement learning step, we were able to generate two candidates: one more targeted towards fluent translations, the other towards more literal but sometimes slightly unnatural translations. In the end we chose to submit the more fluent system as our primary submission.

2 Supervised Fine-Tuning

For supervised fine-tuning (SFT), we begin with the released Gemma 3 27B model. We use parallel data including both human-generated texts as well as synthetic data generated by Gemini (Gemini Team, 2025). In addition we include human-generated Multidimensional Quality Metrics (MQM) translation error annotation data as made available from the WMT evaluation campaigns,[1] as well as generic instruction-following data. We use the public Gemma Kauldron SFT tooling[2] to fine-tune the Gemma 3 27B pretrained checkpoint. For fine-tuning we use the AdaFactor (Shazeer and Stern, 2018) optimizer with a learning rate of 0.0001 and a batch size of 64, running for 20k steps.

2.1 SFT Data

We used 4 different types of data for the Supervised Fine Tuning step.

Synthetic Gemini-Generated Translation Data Our synthetic data is generated using MADLAD-400 as the monolingual source (Kudugunta et al., 2023). The MADLAD-400 sources are first bucketed by length, and then sampled in each bucket to obtain 1 million source segments for each language pair we wish to generate synthetic data for. We then run a preliminary filtering step across these source segments where we take 2 samples from Gemini 2.5 Flash (1 greedy decoding, 1 sampled at temperature=1.0) and compare their scores according to MetricX 24-QE (Juraska et al., 2024). We

[1] https://github.com/google/wmt-mqm-human-evaluation

[2] https://kauldron.readthedocs.io/en/latest/

Proceedings of the Tenth Conference on Machine Translation, pages 723–731
November 8-9, 2025

select the 60k sources where the sample achieves the largest improvement over the greedy decoding. The intuition behind this source filtering approach is that we wish to select sources that will benefit the most from 128-sample QE decoding, so we use 2 samples as a low-cost approximation. We generate at two distinct lengths this way: individual sentences and text blobs of up to 512 tokens. This way we aim to support both translations of individual segments as well as longer texts.

After this selection process, for each of the 60k sources for each language pair we generate 128 samples from Gemini 2.5 Flash and then apply a MetricX 24-QE filter to select the best-performing examples. In order to avoid formatting issues or erroneous translations, we apply an additional filtering step, based again on Gemini 2.5 Flash. This methodology was applied to the language pairs listed in Table 1. First SFT experiments were carried out on a subset of this data, marked in bold in the table.

For translations into Serbian we created a synthetic data variant in both Cyrillic and Latin script with some post-processing filters based on unicode ranges to make sure the translations are in the correct script. The goal of the synthetic data is to cover all languages relevant for the shared task. Except for Bhojpuri, Bengali and Maasai, the data covers all languages of the primary translation task as well as the multilingual subtask: synthetic data generation for Bhojpuri and Bengali did not finish in time for the shared task submission, and we decided to exclude Maasai due to quality concerns given the extremely low-resource nature of the language. In addition, Maasai was not covered by the multilingual pre-training of the MetricX base model, so that QE scores are likely not reliable.

Human-Generated Translation Data To increase the diversity and script coverage of the data we also include data for additional lower-resource languages. For these languages, due to uncertainty about the quality of Gemini-generated synthetic data, we opt to use human-generated parallel data instead. This data comes from the SMOL (Caswell et al., 2025) and GATITOS (Jones et al., 2023) datasets. SMOL covers 221 languages and GATITOS covers 170. This data was only used for the SFT stage, not RL.

Human-Generated MQM Data We include MQM data from WMT 2020 - 2023 (Lommel et al., 2014; Freitag et al., 2021) in the general training data mix. The intention is to increase the diversity of the training data and add information on translation error scoring. The model response is formatted as JSON as seen in Figure 1. A model fine-tuned only on the MQM portion is used as an AutoMQM model (Fernandes et al., 2023) for RL (see Section 3).

Generic Instruction-Following Data Our SFT mixture also includes 40% generic instruction-following data from the original Gemma 3 mixture. The purpose of including this data is to prevent the model from overfitting to the translation task and to maintain generic instruction-following capabilities.

2.2 Translation Performance

In order to measure the performance during the development cycle, we used a subset of the WMT24++ (Deutsch et al., 2025) corpus. We selected those language pairs (starting from English) that are also included in the WMT25 evaluation campaign, i.e. English to Arabic (Egypt), Chinese (Simplified), Czech, Estonian, Icelandic, Italian, Japanese, Korean, Russian, Serbian and Ukrainian.

In Table 2 results for the SFT approach are shown. We first experimented with running SFT on a set of 17 languages, some of which were included in the WMT25 set of languages. On this setup we saw improvements both on MetricX-24-XXL (Juraska et al., 2024) and COMET22 (Rei et al., 2022) (although only slight), but we saw a significant degradation on CHRF (Popović, 2015). Examining the produced translations, we saw a typical case of overfitting: the languages covered by our dataset saw improvements in translation quality, while those not included suffered from important degradations, especially those with alphabets not included in the data (which explained the big drop in CHRF).

With this observation, we designed a setup that tried to balance the improvements while keeping the overall performance. We lowered the learning rate, froze the embeddings and added generic SFT data derived from the Gemma 3 post-training setup, as well as the MQM data. With this setup, we were able to improve the quality as measured with MetricX and COMET22, while also recovering the original CHRF score. We were able to not only avoid drops but even see gains for languages that were not covered by the translation data mix. For example Bengali dropped from 41.83 CHRF

English (en)	↔	**Arabic (ar)**, **Chinese (zh)**, Czech (cs), **Dutch (nl)**, Estonian (et)*, Farsi (fa)*, **French (fr)**, **German (de)**, Greek (el)*, **Hindi (hi)**, **Indonesian (id)**, Indonesian (id)*, Icelandic (is)*, **Italian (it)**, **Japanese (ja)**, Kannada (kn)*, **Korean (ko)**, Lithuanian (lt)*, Marathi (mr)*, **Polish (pl)**, **Portuguese (pt)**, Romanian (ro)*, **Russian (ru)**, Serbian (sr)*, **Spanish (es)**, Swedish (sv)*, **Thai (th)**, **Turkish (tr)**, Ukrainian (uk), **Vietnamese (vi)**
Japanese (ja)	↔	Chinese (zh)

Table 1: List of language pairs for which we generated synthetic data. For language pairs marked with * formatting filtering was not applied. Languages in bold were included in the first set of experiments

System	MetricX	COMET22	CHRF
Baseline (Gemma 3)	3.08	82.7	41.3
SFT on 17 langs	2.94	82.8	37.7
+ general setup	2.81	83.8	41.1
+ WMT25 langs	2.86	84.4	44.2

Table 2: Supervised fine-tuning results on WMT priority languages. "17 langs" refers to the 17 language pairs marked in bold in Table 1.

to 13.87 in the initial SFT setup, while improving to 45.40 in the general setup (and 46.10 when increasing the language coverage). That is despite not being covered by SFT data. This shows the importance of carefully selecting the fine-tuning setup and monitoring languages/scripts outside of the training data.

Lastly we expanded our parallel dataset to all languages included in WMT25 (except Bhojpuri, Maasai and Bengali), which provided an additional boost in COMET22 and CHRF, with a negligible drop in MetricX.

3 Reinforcement Learning

We performed reinforcement learning on top of the SFT checkpoint, using an ensemble of metrics as reward models, to further boost translation quality.

3.1 Reward Models

We used the following metrics as reward models during RL:

- MetricX-24-XXL-QE (Juraska et al., 2024), a learned, regression-based translation metric producing a floating point score between 0 (best) and 25 (worst), matching the standard MQM score range (Freitag et al., 2021). MetricX scores were linearly rescaled, using $5.0 - \texttt{score}$, when computing rewards, so that higher scores indicate better quality. Although MetricX can take source, reference, and hypothesis as input, we passed in an empty reference to use it as a QE score only.

- Gemma-AutoMQM-QE, a finetuned AutoMQM model (Fernandes et al., 2023). This model was initialized from the Gemma3-27B-IT checkpoint (Gemma Team, 2025), and was trained on MQM ratings data from WMT 2020 - WMT 2023 (Lommel et al., 2014; Freitag et al., 2021). Default MQM weights (Freitag et al., 2021) were used in computing (token-level) rewards from AutoMQM outputs. As with MetricX, it ignores the reference translation.

- Generalist reward model covering many tasks, including reasoning, instruction following, and multilingual abilities, adapted from the general Gemma 3 post-training setup (Gemma Team, 2025).

We used RL algorithms extended to support token-level advantages, which were added to the advantages computed from sequence-level rewards. This allowed us to use fine-grained, span-level reward signals from AutoMQM directly, for improved credit assignment and training efficiency in the spirit of Ramos et al. (2024). See Figure 1 for an illustration of how MetricX and AutoMQM rewards were (additively) combined during advantage computation. The combined advantages were then batch-normalized.

3.2 Language Distribution

For RL we used the same translation data as for SFT[3], but ignored the (synthetic) references, since

[3]Except for GATITOS and SMOL, which were used in SFT only.

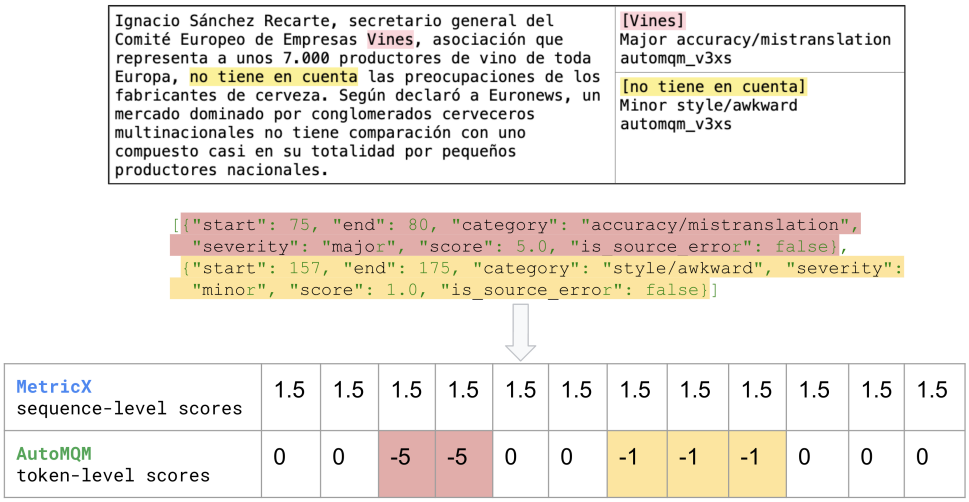
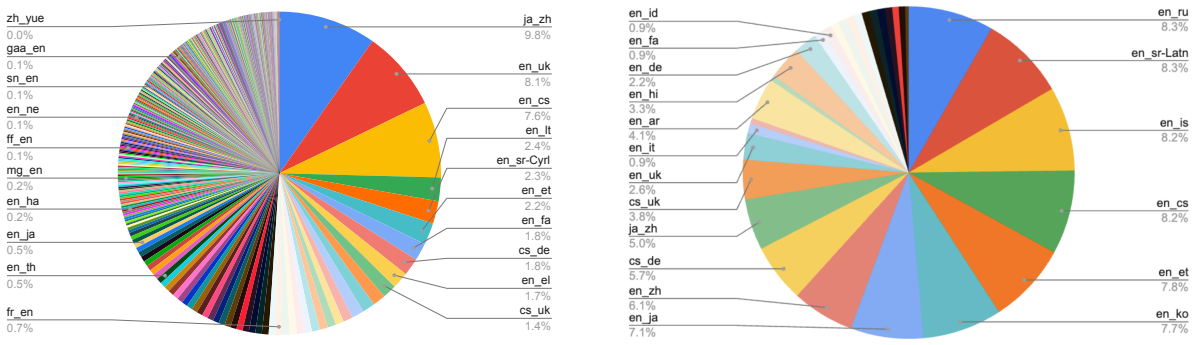


Figure 1: Illustration of how sequence-level and token-level rewards are additively combined during advantage computation in RL. Note that advantage is computed from sequence-level rewards as 'reward-to-go', meaning that rewards are broadcast uniformly to every token.



(a) SFT data mixture.

(b) RL data mixture.

Figure 2: Language distribution (token count) in the GemTrans data mixtures.

the rewards we used were reference-free. Furthermore, we included the WMT languages missing from the SFT step (*bn* and *bho*).

We re-balanced the RL data with UniMax sampling (Chung et al., 2023) to balance the distribution of language pairs in the final mixture, resulting in subsampling the high resource language pairs so that all have the same weight. The final proportion of languages for the SFT and RL phases can be found in Figure 2. While the RL prompt set was (approximately) a subset of the SFT data (note that the RL prompt set excluded the non-translation SFT split), we hypothesized that further gains in performance were still possible, given that the RL learning objective is very different from that of SFT, and improvements from RL (e.g., learning to not hallucinate) should generalize independently of the prompt set used.

System	MetricX	COMET22	CHRF
Baseline (Gemma 3)	3.08	82.7	41.3
+ full SFT	2.86	84.4	44.2
+ RL GEMTRANS1	2.40	83.7	37.4
+ RL GEMTRANS2	2.85	83.7	40.4

Table 3: Reinforcement learning results on WMT priority languages.

3.3 Translation Performance

We experimented with different combinations of reward functions and data conditions on small scale experiments, starting from the fine-tuned model described in Section 2. Our primary submission, "GEMTRANS1", used the ensemble of reward models described in Section 3.1. Our secondary submission, which we refer to as "GEMTRANS2", used

the reference-based version of MetricX-24-XXL (with the same synthetic references used for SFT, as described in Section 2.1), and used a prompted, rather than finetuned, AutoMQM reward model. This prompted AutoMQM model used the same prompt format as the finetuned model. Translation results are shown in Table 3. It can be seen that GEMTRANS1 managed to improve significantly on MetricX, although at the cost of drops in COMET22 and CHRF. GEMTRANS2 was not able to improve over the base SFT system in any metric. However, we still kept it as a candidate for submission, keeping in mind that GEMTRANS1 may potentially be overfitting to MetricX.

4 Automatic Post-editing

In order to minimize formatting errors, we run an additional post-editing pass on the resulting translations. Each model post-edits its own translations, i.e. they serve both as translation and post-edit systems. We prompt the model to just fix the formatting, without altering the text of the translation, using the prompt shown in Appendix A. The examples were chosen from errors we spotted during the development process of the model. On automatic metrics we saw small but consistent improvements for both GEMTRANS1 and GEMTRANS2 systems.

5 Final Submission

As described in Section 3.3, we ended up with two main candidates for submission. MetricX showed a clear preference for GEMTRANS1, although there was a clear drop in CHRF.[4] This might be a signal of the model overfitting to MetricX, so in order to get a better picture, we prompted Gemini to compare a subset of the translations of the WMT25 test sets, assigning a score of +1 if GEMTRANS1 was preferred, and -1 otherwise. The results can be found in Table 4, and show a preference for GEMTRANS1.

Additionally we also prompted Gemini to perform an MQM evaluation with additional quality scoring (similar to this year's setup in the shared task evaluation). In this case, the system preferred GEMTRANS2, both in terms of the quality score as well as MQM. When looking into the decomposition into accuracy and fluency scores, the MQM analysis shows better accuracy for GEMTRANS2, but at the cost of fluency.

[4]We had already discarded the failure cases described in Section 2.2 for this setup.

Language Pair	Mean
cs→de_DE	+0.22
cs→uk_UA	−0.07
en→ar_EG	+0.24
en→bho_IN	+0.23
en→cs_CZ	−0.07
en→et_EE	+0.15
en→is_IS	+0.26
en→ja_JP	+0.26
en→ko_KR	+0.31
en→mas_KE	+0.31
en→ru_RU	+0.26
en→sr_Latn_RS	+0.20
en→uk_UA	+0.31
en→zh_CN	+0.11
ja→zh_CN	+0.12
Average	+0.17

Table 4: Gemini side-by-side scores for the different language pairs. A positive score represents preference for GEMTRANS1, a negative score preference for GEMTRANS2.

Lastly we manually spot checked the translations (focusing on German outputs). We can confirm the conclusions of the MQM analysis, as we found the output of GEMTRANS1 to be more fluent than that of GEMTRANS2. Table 5 shows some examples comparing the two systems. In the first example, GEMTRANS1 uses the informal form "du" instead of the formal "Sie", which is more appropriate for the context of YouTube videos, and in general the flow of the text is more natural than that of GEMTRANS2. The second example shows a more interesting effect. We saw that GEMTRANS1 in several cases adds "explanations" to its output, e.g. it expands the UCL and LSE abbreviations in the example shown in Table 5. In general we found this added information to be useful and appropriate, and we felt that it indeed increased the translation quality. However, in a more strict sense, these could be interpreted as "accuracy" errors, which would explain the error numbers shown in Table 6. GEMTRANS2 on the other hand seemed indeed to be a more accurate model, but it sometimes produced "unnatural" translations.

Of course all that glitters is not gold, and we also found examples where the added text from GEMTRANS1 was incorrect or even hallucinated. As for the question of overfitting to MetricX, we can-

Source	So that's it for today's ikea hacks video. I hope that you enjoy this one and give it a big thumbs up if you did and let me know in the comment section below which one of these projects was your favorite. `[...]`
GEMTRANS1	Das war's für heute mit unserem Video zu IKEA-Hacks. Ich hoffe, es hat **euch** gefallen, und wenn ja, **gebt** uns gerne ein "Daumen hoch"! **Schreibt** uns auch in den Kommentaren, welches dieser Projekte **euch** am besten gefallen hat. `[...]`
GEMTRANS2	Das war's für heute mit diesem Video zu IKEA-Hacks. Ich hoffe, es hat **Ihnen** gefallen, und **geben Sie** ihm einen Daumen hoch, wenn das der Fall war, und **lassen Sie** mich im Kommentarbereich unten wissen, welches dieser Projekte **Ihr** Favorit war. `[...]`

Source	Game 17 of #UniversityChallenge is the first of eight knockout games in the second round. This was UCL and LSE, and UCL gave an outstanding performance: `[...]`
GEMTRANS1	Das 17. Spiel der #UniversityChallenge ist das erste von acht K.O.-Spielen in der zweiten Runde. Es wurde zwischen der UCL **(University College London)** und der LSE **(London School of Economics)** gespielt, und die UCL zeigte eine herausragende Leistung: `[...]`
GEMTRANS2	Spiel 17 von #UniversityChallenge ist das erste von acht K.O.-Spielen in der zweiten Runde. Es handelte sich um UCL und LSE, und UCL zeigte eine herausragende Leistung: `[...]`

Table 5: Translation examples for both systems under consideration. GEMTRANS1 generally produces more fluent translations, sometimes with added explanations that help the understanding of the text.

		MQM			
System	Q	Total	Acc	Fl	O
GEMTRANS1	68.0	15.5	13.6	1.1	0.8
GEMTRANS2	75.6	11.4	8.4	1.9	1.0

Table 6: MQM evaluation by prompting Gemini for MQM and quality (Q) scores. The MQM scores are additionally split into accuracy (Acc), fluency (F) and "other" (O) categories.

not completely discard this hypothesis, but we did not see any evidence of pathological outputs that might be gaming the metric. We decided to move forward with this system, as the general quality indeed seemed to be superior to that of GEMTRANS2, and this constituted our primary submission for the shared task.

6 Findings of the Human Evaluation

The organizers of WMT shared the results of the human evaluation. The performance of GEMTRANS1

Language Pair	Cluster	Rank
cs→de_DE	2	9-14
cs→uk_UA	2	4-8
en→ar_EG	9	11-14
en→bho_IN	n/a	n/a
en→cs_CZ	4	13-16
en→et_EE	7	10-12
en→is_IS	9	12-12
en→it_IT	1	1-4
en→ja_JP	3	12-16
en→ko_KR	2	5-10
en→mas_KE	n/a	n/a
en→ru_RU	3	13-16
en→sr_Cyr_RS	6	8-9
en→uk_UA	2	4-8
en→zh_CN	2	5-10
ja→zh_CN	5	14-15

Table 7: Human evaluation results for the GEMTRANS1 system.

has been summarized in Table 7. It can be seen that GemTrans performs generally in the top 3 clusters for 8 of the 14 language pairs where it was evaluated by humans, being in the first one for English to Italian. The first clusters are usually taken by large systems with a much bigger parameter count. The ranks show a wider variance, due to the highly competitive landscape of this year's shared task and the big number of participating systems.

For English to Arabic GEMTRANS1 obtained a very low human score, along with all other systems in its cluster. This was the result of the system failing to produce the correct Arabic dialect (Egyptian). In a related fashion, Bhojpuri showed low automatic scores (thus GEMTRANS1 was not included in the human evaluation) and we suspect that GEMTRANS1 failed to generate Bhojpuri, falling back to Hindi instead.

In the report about the shared task, in the additional analysis of the Serbian translations, the organizers explicitly highlight the GEMTRANS1 system for a "notable amount of idiomatic translations, even more than humans" (Kocmi et al., 2025), although they also point out a "relatively high number of errors". These findings agree with our own observations about the system.

7 Conclusions

We have presented the Google Translate Research submission to the WMT25 evaluation campaign. Starting from Gemma 3 we used supervised fine-tuning and RL to boost translation performance, in addition to a small automatic post-editing step to improve the formatting of the translations.

Out of the two final candidate systems, we found that one was more tailored towards fluency while the other one was more tailored towards accuracy (possibly illustrating the tradeoff discussed by e.g. Flamich et al. (2025); Schleiermacher (1816); Dryden (1685)). After evaluation with automatic metrics as well as manual inspection, we decided to move forward with the more fluent system, as it seemed to produce generally higher quality translations.

Whether that was the correct decision may largely depend on the criterion of the human evaluators. Quality of machine translation is indeed in the eye of the beholder, and the more "free-style" translations produced by the system may not be the preferred ones if more literal translation are desired, closer to the source sentence. This dichotomy again highlights the difficulties of machine translation evaluation, and may be indeed point towards new research directions (for both MT generation and evaluation), where the intent of the translation can play a more relevant role.

References

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, et al. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. *arXiv preprint arXiv:2502.12301*.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, et al. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

John Dryden. 1685. *Sylvae [Translator's preface]*. A Scolar Press facsimile. Scolar Press.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. You cannot feed two birds with one score: the accuracy-naturalness tradeoff in translation.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gemini Team, Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. 2025. Gemma 3 technical report.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Miguel Moura Ramos, Tomás Almeida, Daniel Vareta, Filipe Azevedo, Sweta Agrawal, Patrick Fernandes, and André FT Martins. 2024. Fine-grained reward optimization for machine translation using error severity mappings. *arXiv preprint arXiv:2411.05986*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Friedrich Schleiermacher. 1816. *Über die verschiedenen Methoden des Übersetzens*. Abhandlungen der Königlichen Akademie der Wissenschaften in Berlin. Walter de Gruyter GmbH.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Appendix

A APE Prompt

```
You are a copy-editor fixing formatting issues related to translation. You will see
a source in {src_lang} and its translation in a {tgt_lang}. If there are formatting
errors, fix them. Here are examples of things to fix:

{examples}

Do NOT fix the translations themselves. Only fix the sorts of minor errors described
above, IF they exist. MOST TRANSLATIONS WILL NOT NEED ANY CORRECTIONS!

Only output the fixed translation, with no additional formatting or chattiness. Here
is the example to fix:

source={src}
output={out}
corrected=
```

Examples

- Mismatching quotation marks:

  ```
  source="Let her go!"
  output="¡Déjala ir!"
  corrected="¡Déjala ir!"
  ```

- "user" omitted from beginning:

  ```
  source=@user38 heard this essay was good and it is
  output=Ich habe gehört, dieser Artikel ist gut, und das ist er auch.
  corrected=@user38: Ich habe gehört, dieser Artikel ist gut, und das ist er auch.
  ```

- HTML tag translated, instead of preserved as tag

  ```
  source=<contents for=sec2>section 2...</contents>
  output=<Inhalt für=sec2>Abschnitt 2...</Inhalt>
  corrected=<contents for=sec2>Abschnitt 2...</contents
  ```

DLUT and GTCOM's Large Language Model Based Translation System for WMT25

Hao Zong[1,2] **Chao Bei**[2] **Conghu Yuan**[2]
Wentao Chen[2] **Huan Liu**[2] **Degen Huang**[1]*
[1]Dalian University of Technology
[2]Global Tone Communication Technology Co., Ltd.
zonghao@mail.dlut.edu.cn
{beichao, yuanconghu, chenwentao and liuhuan}@gtcom.com.cn
huangdg@dlut.edu.cn

Abstract

This paper presents the submission from Dalian University of Technology (DLUT) and Global Tone Communication Technology Co., Ltd. (GTCOM) to the WMT25 General Machine Translation Task. Amidst the paradigm shift from specialized encoder-decoder models to general-purpose Large Language Models (LLMs), this work conducts a systematic comparison of both approaches across five language pairs. For traditional Neural Machine Translation (NMT), we build strong baselines using deep Transformer architectures enhanced with data augmentation. For the LLM paradigm, we explore zero-shot performance and two distinct supervised fine-tuning (SFT) strategies: *direct translation* and *translation refinement*. Our key findings reveal a significant discrepancy between lexical and semantic evaluation metrics: while strong NMT systems remain competitive in BLEU scores, fine-tuned LLMs demonstrate marked superiority in semantic fidelity as measured by COMET. Furthermore, we find that fine-tuning LLMs for direct translation is more effective than for refinement, suggesting that teaching the core task directly is preferable to correcting baseline outputs.

1 Introduction

The field of machine translation is undergoing a profound paradigm shift, marked by the ascent of general-purpose Large Language Models (LLMs) that challenge the dominance of specialized encoder-decoder Neural Machine Translation (NMT) architectures (Vaswani et al., 2017). For years, NMT systems, meticulously trained on vast parallel corpora, have been honed into highly effective, specialized tools for a single task: translation (Ott et al., 2019). In contrast, LLMs, pretrained on web-scale multilingual and multimodal data, emerge as powerful generalists, possessing not only cross-lingual capabilities but also extensive world knowledge and reasoning skills (Brown et al., 2020), which they can apply to translation with remarkable zero-shot proficiency. This dichotomy between the "specialized artisan" (NMT) and the "generalist polymath" (LLM) raises critical questions about the future trajectory of machine translation research.

This transition is further complicated by an evolution in evaluation philosophy. The community is increasingly moving away from lexical overlap metrics like BLEU (Papineni et al., 2002), which may unduly penalize valid, fluent translations that diverge stylistically from a single reference. The rise of semantic-aware metrics such as COMET (Rei et al., 2020) and its successor, XCOMET-XL (Guerreiro et al., 2023), reflects a demand for evaluations that prioritize meaning and fidelity. This shift is particularly pertinent when comparing NMT and LLMs, as LLMs often excel at producing semantically coherent and contextually appropriate outputs that might be lexically dissimilar to the reference. A core challenge, therefore, is to conduct a fair comparison that accounts for this evaluation dichotomy.

In this paper, we leverage our participation in the WMT25 General Machine Translation task as a standardized testbed to systematically investigate this ongoing paradigm shift. Our work is guided by two central research questions (RQs):

1. **(RQ1)** How do the performance characteristics of specialized NMT systems and general-purpose LLMs diverge, particularly under the contrasting lenses of lexical (BLEU) and semantic (COMET) evaluation metrics?

2. **(RQ2)** Among supervised fine-tuning (SFT) strategies for adapting LLMs to translation, which is more effective: direct instruction on source-to-target mapping (*direct translation*), or training the model to correct outputs from a baseline system (*translation refinement*)?

*Corresponding Author

Proceedings of the Tenth Conference on Machine Translation, pages 732–739
November 8-9, 2025

To address these questions, we developed a comprehensive suite of systems. Our NMT pipeline features deep Transformer models trained with the `fairseq` toolkit, enhanced by data augmentation. Our LLM pipeline is built upon the powerful Gemma3 model family (Team et al., 2025), which we adapt using the `LLaMa-Factory` framework (Zheng et al., 2024). Our main contributions are: (1) a robust empirical comparison of NMT and LLM systems across five language pairs, revealing a significant divergence between lexical and semantic evaluation scores; (2) a direct analysis of two distinct LLM fine-tuning strategies, demonstrating the superior efficacy of direct translation; and (3) insights into the qualitative differences between the outputs of these systems, highlighting the semantic strengths of modern LLMs.

2 Task Description

The core of this task is bilingual text translation. The data, sourced using the 'mtdata' tool(Gowda et al., 2021) from the official WMT25 repository, consists of both parallel and monolingual corpora. Table 1 provides a detailed breakdown of the training data statistics. For our development and testing sets, we used newstest2019 for the Czech→German direction, wmttest2024 for Czech→Ukrainian, English→German, and English→Ukrainian, and flores200-devtest (NLLB Team, 2022) for English→Serbian.

Data Type	Number of Sentences
Parallel Data	
cs-de	120.39M
cs-uk	10.62M
en-uk	24.6M
en-ru	77.5M
en-sr	114.04M
Monolingual Data	
English	35M
Czech	42.6M
Ukrainian	14.8M
German	72.8M
Serbian	56.8M
Russian	56.2M
Development Sets	
cs-de	1997
cs-uk	2316
en-uk	997
en-ru	997
en-sr	1012

Table 1: Statistics for the training and development datasets.

3 Methodology

Our methodology is designed as a comparative study of two distinct translation paradigms. We first establish a strong baseline representing specialized NMT systems and then build upon a generalist LLM foundation, exploring different adaptation strategies.

3.1 Data Foundation: Preprocessing and Quality Filtering

A high-quality dataset is the bedrock of any translation system. Our preprocessing pipeline is standardized across all languages and includes punctuation normalization, tokenization, Truecasing, and Byte Pair Encoding (BPE) (Sennrich et al., 2015) to manage vocabulary size and handle rare words.

Beyond standard preprocessing, we implemented a rigorous quality filtering stage using the CometKiwi tool ('wmt23-cometkiwi-da-xl' model) (Rei et al., 2023). For our LLM fine-tuning, we adopted a nuanced data selection strategy. Rather than simply taking the top-N scoring sentence pairs, we extracted a 100,000-pair subset ranked between the 10,000th and 110,000th positions. This decision is based on the hypothesis that the absolute highest-scoring pairs often consist of overly simplistic, short, or formulaic sentences (e.g., from translation memories), which can lead to models that are fluent but lack complexity. By targeting a "high-quality but challenging" segment, we aim to create a more diverse and robust instruction dataset for fine-tuning.

3.2 Paradigm 1: Specialized NMT Systems

To represent the best in specialized NMT, we employed a deep Transformer architecture ('transformer_wmt_en_de' configuration in `fairseq`). These models, featuring 24 encoder and 24 decoder layers, serve as our high-performance baseline. To maximize the utility of available monolingual data—a cornerstone of competitive NMT—we incorporated iterative data augmentation:

1. **Back-Translation (BT):** We trained reverse-direction models (e.g., ru→en) to translate target-language monolingual data into the

source language, creating a large, synthetic parallel corpus to augment the primary training data.

2. **Forward-Translation (FT):** The improved models from the BT step were then used to translate source-language monolingual data, further enriching the training mixture in a subsequent iteration.

3.3 Paradigm 2: Generalist LLM-based Systems

Our exploration of the generalist paradigm centers on the Gemma3 model family, selected for its strong preliminary multilingual performance. Our approach systematically moves from zero-shot evaluation to targeted adaptation.

3.3.1 Foundation Model and Zero-Shot Baseline

We first established a zero-shot baseline by evaluating several prominent instruction-tuned LLMs (including the Qwen3 and Gemma3 series) on our development sets using a direct translation prompt. This step measures the intrinsic, out-of-the-box translation capabilities of these models without any task-specific training.

3.3.2 Supervised Fine-Tuning (SFT) Strategies

To adapt Gemma3 for high-quality translation, we investigated two distinct SFT strategies, each testing a different hypothesis about how LLMs best learn this complex task. **To achieve the most thorough adaptation possible, we performed full-parameter fine-tuning, allowing all weights of the base model to be updated during the training process.** This approach, while computationally intensive, ensures that the model can fully specialize its internal representations for the translation task.

Strategy 1: Direct Translation. In contrast, this strategy reframes the task from generation to a more complex process of critique and correction. The model is provided with a triplet: the source text, a potentially flawed translation from our NMT baseline, and the high-quality reference. The hypothesis is that by learning to identify and correct errors—essentially, learning the "delta" between a mediocre and an excellent translation—the model develops a more nuanced understanding of quality, error patterns, and stylistic appropriateness. This task is guided by a prompt that casts the LLM in the role of a professional "post-editor":

```
You are an expert in {Source
    language}-{Target language}
    translation, with a deep
    understanding of both languages'
    cultural nuances. Your
    translations are accurate, fluent
    , and elegant. Please translate
    the following {Source language}
    text into {Target language}. Only
     output the translation.

{Source language} text: {Source text
    }
{Target language} translation:
{Target text}
```

Strategy 2: Translation Refinement. This strategy reframes the task from generation to critique and correction. The model is provided with a source text, a potentially flawed translation from our NMT baseline, and the high-quality reference. It is then instructed to "polish" or "refine" the baseline translation. The hypothesis here is that learning to identify and correct errors is a more cognitively demanding task that could foster a deeper, more nuanced understanding of translation quality, error patterns, and stylistic appropriateness, potentially leading to a more robust translator. The prompt for translation refinement is as follows:

```
You are a professional {Source
    language}-{Target language}
    translation refinement expert who
     excels at making machine-
    translated content more natural
    and fluent, ensuring it aligns
    better with the target language's
     norms and contexts. Based on the
     provided source text and machine
     translation, refine and modify
    the translation to make it more
    accurate and natural.

{Source language} source: {Source
    text}
{Target language} translation: {
    Baseline translation text}
{Target language} corrected
    translation:
{Target text}
```

4 Experimental Setup

Our experiments were designed to ensure a fair and reproducible comparison between the NMT and LLM paradigms.

4.1 NMT System Configuration

We used the `fairseq-py` toolkit for all NMT experiments. Our deep Transformer models were trained with the following configuration:

- **Architecture:** 'transformer_wmt_en_de' (24 encoder/decoder layers, 16 attention heads, embedding size of 1024).
- **Optimizer:** Adam (Kingma and Ba, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.98$.
- **Learning Rate Schedule:** Inverse square root scheduler with a warm-up of 4,000 steps and a peak learning rate of 5×10^{-4}.
- **Regularization:** Dropout was set to 0.3 for the attention and activation functions, and label smoothing of 0.1 was applied.
- **Batching:** We used a maximum of 4096 tokens per batch per GPU. Models were trained for 100,000 steps or until convergence on the development set.

4.2 LLM System Configuration

All LLM experiments were conducted using the `LLaMa-Factory` framework.

- **Base Models:** We used the instruction-tuned versions of the Gemma3 family: 'Gemma3-12B-it' and 'Gemma3-27B-it'.
- **Fine-Tuning Method:** We employed **full-parameter supervised fine-tuning**. This involves updating all of the model's weights, rather than using a parameter-efficient method.
- **Hyperparameters:** Models were trained for 3 epochs over the 100k-pair instruction dataset. We used the AdamW optimizer with a learning rate of 2×10^{-5}, a cosine learning rate scheduler, and a warm-up ratio of 0.03. Training was performed with `bfloat16` mixed-precision to optimize memory usage and throughput.

4.3 Evaluation Metrics

To provide a multifaceted view of translation quality, we report scores from two distinct metrics:

- **sacreBLEU** (Post, 2018): A standardized implementation of BLEU that measures n-gram precision against a reference translation. It primarily reflects lexical similarity.
- **XCOMET-XL** (Guerreiro et al., 2023): A state-of-the-art semantic metric that uses a large pre-trained model to assess the meaning equivalence between the source, hypothesis, and reference. This metric aligns more closely with human judgments of translation quality.

5 Results and Discussion

In this section, we analyze our experimental results to answer the research questions posed in the introduction. We dissect the performance of each paradigm and discuss the implications of our findings. The comprehensive results for our NMT baselines and zero-shot LLM evaluations are consolidated in Table 2.

5.1 RQ1: NMT vs. LLMs and the BLEU-COMET Dichotomy

Our first research question explores the performance divergence between specialized NMT and generalist LLMs. The results in Table 2 reveal a fascinating and consistent trend that we term the BLEU-COMET dichotomy.

NMT systems remain formidable competitors on lexical metrics, often outperforming even large LLMs in BLEU score. This is most evident in the en→sr direction, where the NMT baseline achieves a BLEU score of 38.44, significantly higher than any other system. Similarly, for cs→de, the NMT baseline's BLEU of 29.67 is the highest in its category. Regarding data augmentation, back-translation shows a clear benefit for lower-resource pairs (e.g., providing a 1.73 BLEU point gain for cs→uk), but its impact diminishes or even slightly degrades BLEU on high-resource pairs like cs→de. Furthermore, forward-translation consistently proves detrimental to performance across most pairs, likely due to the introduction of unmitigated noise.

In contrast, LLMs exhibit a clear and striking superiority in semantic fidelity, even in a zero-shot setting. The Gemma3-27B-it model achieves the highest zero-shot COMET score in four out of five language pairs. The most dramatic example is en-uk, where the Gemma3-27B-it's COMET score of 80.31 massively surpasses the NMT system's 67.55, despite having only a marginal advantage in BLEU. This pattern holds for cs→de (94.24 vs. 93.71), cs→uk (91.47 vs. 88.65), and en→ru (91.73 vs. 91.55). The only exception is en→sr, where the NMT baseline's COMET score

Model / System	cs→de		cs→uk		en→uk		en→ru		en→sr	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Specialized NMT Systems										
NMT Baseline	**29.67**	93.71	26.59	85.71	25.97	66.02	28.21	90.11	**38.44**	**90.45**
+ Back-translation	29.59	92.80	28.32	88.65	26.22	67.55	**28.86**	91.55	36.05	86.21
+ Forward-translation	26.37	82.50	24.50	82.50	25.25	66.10	28.55	90.81	31.66	83.54
Generalist LLMs (Zero-shot)										
Qwen3-8B	12.72	90.70	13.52	84.25	17.42	67.53	19.79	86.21	10.41	69.25
Qwen3-14B	22.24	92.63	26.57	87.65	23.44	71.52	27.59	87.69	10.58	78.27
Gemma3-12B-it	24.29	93.62	29.24	91.01	25.65	79.08	26.43	91.01	11.42	86.27
Gemma3-27B-it	25.53	**94.24**	**30.50**	**91.47**	**27.22**	**80.31**	28.02	**91.73**	26.62	90.25

Table 2: Comprehensive results comparing our specialized NMT systems against zero-shot performance of generalist LLMs across all five language pairs. While NMT+BT often leads in BLEU, the Gemma3-27B-it model consistently achieves the highest COMET scores, highlighting the BLEU-COMET dichotomy.

Model and SFT Strategy	cs→de		cs→uk		en→uk		en→ru		en→sr	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Gemma3-12B-it Fine-tuned										
Direct Translation SFT	25.45	94.01	27.01	91.23	28.25	80.23	28.31	90.11	31.28	87.16
Refinement SFT	25.71	94.17	26.60	91.57	24.98	79.88	28.01	87.66	30.49	86.59
Gemma3-27B-it Fine-tuned										
Direct Translation SFT	**26.69**	**94.50**	**30.50**	**92.10**	**29.57**	**81.56**	**29.53**	**91.50**	**31.90**	**90.97**
Refinement SFT	26.32	94.32	29.94	91.01	27.83	80.61	29.26	90.50	31.33	88.97

Table 3: Results of supervised fine-tuning on Gemma3 models. The Direct Translation strategy consistently outperforms the Refinement strategy across nearly all models and language pairs. The fine-tuned Gemma3-27B-it with Direct SFT emerged as our best overall system.

is competitive (90.45 vs. 90.25). This powerful trend suggests that LLMs' vast world knowledge allows them to generate more fluent and semantically equivalent translations, a quality that is rewarded by COMET but can be unfairly penalized by BLEU's rigid lexical matching.

5.2 RQ2: Efficacy of SFT Strategies

Our second research question investigates the more effective SFT strategy for adapting LLMs to translation. The results from our fine-tuning experiments, presented in Table 3, provide a decisive answer.

Direct Translation consistently and significantly outperforms Translation Refinement. For both the 12B and 27B model sizes and across all five language pairs, the models fine-tuned with the direct translation task achieved superior scores on both BLEU and COMET. For instance, in the en-uk direction, the Gemma3-27B-it model fine-tuned for direct translation achieved a COMET score of 81.56, while the refinement-tuned model scored only 80.61. Similarly, for en-sr, the direct translation model achieved a COMET of 90.97, a full two points higher than the refinement model's 88.97. The Gemma3-27B-it with Direct Translation SFT emerged as our best overall system, achieving the highest COMET score across the board.

We attribute this clear victory to two primary factors. First, the refinement task introduces a higher cognitive load: the model must simultaneously comprehend the source, analyze the errors in a flawed translation, and generate a correction. This may represent a less direct and noisier learning signal. Second, the provided baseline translation may act as a negative anchor, implicitly constraining the model's output space and preventing it from generating a truly novel and superior translation from scratch. It learns to "edit" rather than to "create."

5.3 Overall Performance and Future Outlook

Our best-performing systems for all language pairs were the Gemma3-27B-it models fine-tuned using the direct translation strategy. As shown by our final official scores in Table 3, these systems achieved competitive results. However, a gap remains when compared to the top-ranking teams in the official evaluation.

Our analysis suggests that while full-parameter SFT on a high-quality 100k dataset is effective, it represents only the initial stage of true model alignment. To reach the highest echelons of translation quality, future work should focus on more

advanced alignment techniques that have proven successful in general-domain LLMs. Promising directions include:

- **Continual Pre-training:** Further adapting the base LLM on large-scale, in-domain monolingual and bilingual data before the SFT stage.
- **Preference Optimization:** Moving beyond standard SFT to methods like Direct Preference Optimization (DPO) (Rafailov et al., 2024), which learns from human or AI-judged preferences between translation candidates, thereby optimizing directly for perceived quality.

This work confirms that while the era of LLMs is here, achieving state-of-the-art translation performance requires more than just scale; it demands sophisticated and targeted adaptation strategies.

6 Conclusion

In this paper, we presented a systematic comparison between specialized Neural Machine Translation (NMT) systems and general-purpose Large Language Models (LLMs) within the framework of the WMT25 General MT Task. Our work was designed to investigate the ongoing paradigm shift in the field, focusing on the divergence in performance characteristics and the efficacy of different LLM adaptation strategies.

Our investigation yielded clear answers to our initial research questions. **First (RQ1)**, we identified a significant and consistent "BLEU-COMET dichotomy." While our highly optimized NMT systems remained competitive, and occasionally superior, in terms of lexical similarity (BLEU), LLMs demonstrated a marked advantage in semantic fidelity (COMET), even in a zero-shot setting. This finding underscores the limitations of traditional metrics in the age of LLMs and highlights the unique ability of these large models to produce fluent and semantically equivalent translations.

Second (RQ2), our experiments on supervised fine-tuning strategies provided a decisive result: direct translation proved to be a more effective adaptation method than translation refinement. We hypothesize that teaching the model the core source-to-target mapping task directly provides a cleaner and more potent learning signal than asking it to perform the more complex, multi-step task of identifying and correcting errors from a baseline system.

Our best systems, based on full-parameter fine-tuning of the Gemma3-27B-it model, achieved highly competitive results. However, our analysis suggests that the next frontier for LLM-based translation lies beyond standard SFT. The path to state-of-the-art performance will require more sophisticated alignment techniques that can better bridge the gap between the LLMs' vast generative capabilities and the nuanced preferences of human evaluation. Future work should therefore prioritize the exploration of methods such as continual pre-training on in-domain corpora and, most promisingly, preference optimization techniques like DPO.

Acknowledgments

This work also receives substantial support from the 2030 Artificial Intelligence Research Institute of Global Tone Communication Technology Co., Ltd. This paper is also supported by National Key Research and Development Program of China (2022ZD0116100).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula

Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma.

2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Yandex Submission to the WMT25 General Translation Task

Nikolay Karpachev Ekaterina Enikeeva Dmitry Popov
Arsenii Bulgakov Daniil Panteleev Dmitrii Ulianov Artem Kryukov Artem Mekhraliev

Yandex

Abstract

This paper describes Yandex submission to the WMT25 General Machine Translation task. We participate in English-to-Russian translation direction and propose a purely LLM-based translation model. Our training procedure comprises a training pipeline of several stages built upon YandexGPT, an in-house general-purpose LLM. In particular, firstly, we employ continual pretraining (post-pretrain) for MT task for initial adaptation to multilinguality and translation. Subsequently, we use SFT on parallel document-level corpus in the form of P-Tuning. Following SFT, we propose a novel alignment scheme of two stages, the first one being a curriculum learning with difficulty schedule and a second one - training the model for tag preservation and error correction with human post-edits as training samples. Our model achieves results comparable to human reference translations on multiple domains.

1 Introduction

We participate in the WMT25 General Machine Translation task and propose a purely LLM-based translation system.

Large Language Models (LLMs) have recently redefined the state-of-the-art in machine translation, demonstrating strong capabilities that yield near-human quality outputs on vast collection of language pairs. Their performance has consistently surpassed that of the previous generation of specialized Neural Machine Translation (NMT) systems, marking a significant paradigm shift in the field. The recent WMT24 General Machine Translation contest provides compelling evidence of this trend, where top-performing systems, predominantly based on LLMs, achieved translation scores remarkably close to human reference translations.

Still, the human parity claim remains disputable and thorough evaluations have shown that for a variety of high-resource morphologically rich languages, although quite high, the performance of LLMs still lags behind professional human translations. All state-of-the-art LLM systems exhibit a noticeable pattern of literal translations with fluency and naturalness of generations significantly worse than that of native human translations.

In this work, we propose a novel pipeline for LLM adaptation to the MT task, building upon our previous year submission (Elshin et al., 2024). The main goal of this work is to explore tools and techniques for improving the performance of an already capable MT-specific LLM model with near human performance.

The system comprises a pipeline of several adaptation stages built upon 7-billion YandexGPT, an in-house proprietary general-purpose language model.

- First, we employ continual pretraining for robust adaptation of general-purpose pretrained model to the task of translation and multilinguality (post-pretrain)
- Following post-pretrain, the model is fine-tuned on a cleaner corpus of automatically collected books translations
- Subsequently, the system undergoes an alignment procedure consisting of two primary stages
  - Contrastive Preference Optimization (CPO) with curriculum learning for low-resource document-level adaptation, targeted at fluency and cohesion improvement
  - Second alignment stage focused on tackling the model shortcomings and tag preservation training

The resulting model (Yandex) was subsequently fine-tuned on the WMT dataset, producing the Yandex+WMT model which constitutes our submis-

Proceedings of the Tenth Conference on Machine Translation, pages 740–752
November 8-9, 2025

sion to the WMT25 General Machine Translation task.

For the English-to-Russian translation direction our resulting system is significantly better than all of the previous year contenders and achieves results comparable to major foundational LLMs on this year's benchmark.

2 System Overview

In this section we describe the training pipeline of the system and details of the inference procedure. We also provide automatic metrics and human evaluation results for the key components of the system.

2.1 Pretrain

The base model that we use is a 7 billion parameter version of YandexGPT, an in-house general-purpose large language model. It is a decoder-only model of an architecture similar to Touvron et al. (2023) trained on a collection of data primarily consisting of Russian and English texts.

In our experiments we use the pre-train stage of YandexGPT as the starting point for machine translation specific fine-tuning.

2.2 Post-pretrain

As an initial adaptation for multilinguality and translation task, we perform a continual pretraining using recipe similar to Alves et al. (2024).

We fine-tune the model using full weights fine-tuning with a standard cross-entropy loss on a mixture of pretrain dataset and MT data with a ratio of translation data of 30%.

The parallel translation data is collected using a matching pipeline similar to Bitextor (Esplà-Gomis, 2009). It involves matching multilingual websites as candidates for parallel documents, followed by a series of alignment and filtering steps (Thompson (2019), Artetxe and Schwenk (2019)).

For each example of parallel translation data, we construct two samples for continual pretraining via concatenation in both ordering variants: english text concatenated with russian translation via two new-line separators and vice-versa.

In our experiments we have observed that the optimal ratio of incorporating MT data is around 30% and it is beneficial to mirror all en-ru training samples in reverse direction.

2.3 Supervised Fine-tuning

Following post-pretrain, we employ supervised fine-tuning (SFT) in order to focus the model solely on the translation task and enforce more precise outputs without hallucinations.

We use an in-house dataset of parallel English and Russian books aligned at the paragraph-level. We apply filtering by length and train on fragments with maximum length of 1k sentence-piece tokens (Kudo and Richardson, 2018). In addition to that, we only use paragraph pairs with the same number of sentences in the source and translation text.

In terms of the training procedure, we have experimented with both standard SFT finetuning and sparse methods like LoRa (Hu et al., 2021) and PTune variants (Liu et al. (2021b), Liu et al. (2021a)).

Our experiments have shown that not only does not the Full Finetuning improve quality upon parameter-efficient training strategies but it also leads to the quality degradation after subsequent alignment. We hypothesize that the root cause for this phenomenon is the "knowledge forgetting". Specifically, more extensive fine-tuning methods make SFT checkpoint less sensitive to the pre-trained LLM knowledge and hence over-optimize for the parallel dataset during the SFT stage.

The resulting system involves training with P-Tuning v2 (Liu et al., 2021a) with two trainable P-Tuning blocks each having the size of 100 ptune tokens placed

- At the start of the input string, before the English source
- Between English source and Russian translation to be generated

In Table 1 we report automatic evaluation results of models from pretrain and SFT stages. We measure MetricX-24 (Juraska et al., 2024) in both reference-based and reference-less QE variants as well as fluency score, which is an in-house monolingual classifier that measures the grammatical and lexical correctness of the Russian translation. We use the same subset of WMT24 test data as in human evaluations for results consistency.

2.4 First-Stage Alignment

In this section we describe the key components of the subsequent step in the training pipeline - the first stage of alignment fine-tuning.

The primary goal of alignment is to effectively make use of another form of training signal - user

Model	METRICX-24-XXL	METRICX-24-XXL-QE	Fluency
YandexGPT Pretrain	4.746	4.262	0.754
MT Postpretrain	4.899	4.499	0.772
SFT (PTune)	4.272	3.854	0.795

Table 1: Comparison of model performance on WMT24 testset.

preferences data or algorithmic rankings of translations with varying quality. In contrast to SFT, alignment techniques like reinforcement learning or contrastive learning allow to perform not only "likelihood" traning on good reference, but also "unlikelihood" training on the data with proven deficiencies or more sophisticated ranking-based approaches.

2.4.1 Sentence-Level Data

The first portion of the training dataset is an internal collection of historic human evaluations of various model generations. This dataset consists of sentence triplets ("source", "winner_translation", "loser_translation") and has a size of several tens of thousand examples.

Namely, "winner_translation" is a preferred hypothesis, and "loser_translation" is a dispreferred. All rankings were done by professional human annotators via a platform similar to Amazon Mechanical Turk.

The samples themselves are single sentences of varying length, typically no more than 100 tokens.

2.4.2 Document-Level Data

As training solely on sentences would not expose the model to complex and practically challenging discourse phenomena of translation like deixis, ellipsis and lexical cohesion we also collected several specific datasets to emphasize such phenomena during training.

The first one is specifically targeted on fluency and coherence improvement (**fluency repair**):

1. Generate paragraph translations with the sentence-level translation model.

2. "Improve" the non-coherent paragraph translations using a monolingual general-purpose LLM.

3. Train on such fluency corrections using contrastive learning, wherein the "positive" hypothesis would be the smoothed and fluent translation and the "negative" hypothesis would be the original sentence-wise translation.

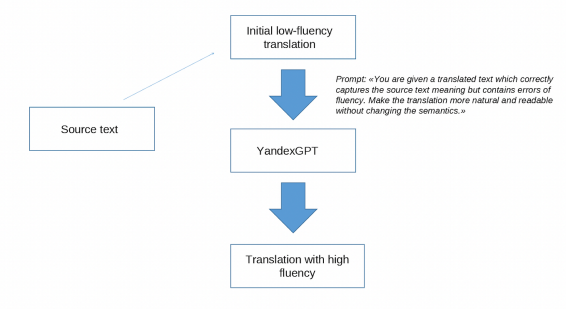


Figure 1: Fluency repair procedure.

We also collect several thousands of side-by-side comparisons of different model generations on paragraph-level source data.

In addition to the contrastive learning triplet data, we found it beneficial to mix triplet data with a small high-quality SFT set of manually written translations created by experts with high language proficiency. We add several thousand such samples to the alignment stage.

2.4.3 Curriculum Learning

In the previous two sections we have outlined the data collection procedure of two parts: sentence-level and document-level data. Sentence-level dataset is significantly larger, consisting of more than 100.000 samples, while the whole document or paragraph-level translations corpus is almost an order less.

This leads to a data imbalance issue during training. Uniform mixture of both sentence- and document-level sources would be highly skewed towards sentence part and in our experiments it produced results highly similar to training only on sentences. Upsampling of document-level data would result in overfitting.

This outlined problem could be handled through training with cirruculum learning (Bengio et al., 2009). We employ a difficulty schedule, first training solely on sentence corpus and then switching to documents at the end of the training.

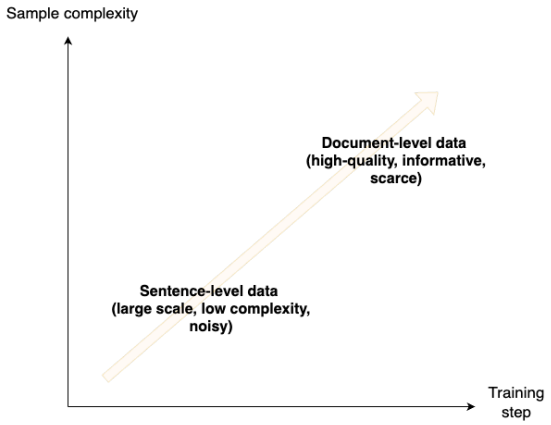


Figure 2: Curriculum learning with sentences-to-documents adaptation.

2.4.4 Training Procedure

We train using Contrastive Preference Optimization (CPO) objective (Xu et al., 2024)

$$\mathcal{L}(\pi_\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Bigg[\log\sigma\Big(\beta\log\pi_\theta(y_w|x) - \beta\log\pi_\theta(y_l|x)\Big)\Bigg]$$

with a cross-entropy regularizer:

$$\min_\theta \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{prefer}} - \underbrace{\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log\pi_\theta(y_w|x)]}_{\mathcal{L}_{NLL}}.$$

We have observed that using higher weights before regularizer cross-entropy term leads to more literal and adequacy-boosting translations (close to the SFT fine-tuning), while giving more weight to the contrastive learning term makes the model much more fluent, but prone to hallucinations.

Overall, the training objective for contrastive triplets is CPO with cross-entropy regulatization weight set to 0.1:

$$\min_\theta \underbrace{\mathcal{L}(\pi_\theta, U)}_{\mathcal{L}_{prefer}} - 0.1 \cdot \underbrace{\mathbb{E}_{(x,y_w)\sim\mathcal{D}}[\log\pi_\theta(y_w|x)]}_{\mathcal{L}_{NLL}}.$$

For the high-quality references, we use standard SFT with learning rate 20 times higher than those for contrastive samples.

We train with one epoch and follow a triangular learning rate schedule with the warmup of 10% of steps and linear decay.

Sentence- and document-level portions of the data are shuffled and the document-level data consists of fluency repair and side-by-side triplets mixed with SFT samples for high-quality references.

2.5 Second-Stage Alignment

During the second stage of alignment, our primary goal is to precisely address specific model deficiencies while developing structure-preserving capabilities. This phase focuses on fine-grained tuning of the model through targeted interventions.

We concentrate on two parallel objectives: first, we aim to fix the common errors of the model after the initial alignment phase. Second, we enhance the model's ability to maintain specific structural elements of the input data (such as HTML tags, list orderings, dialogue formatting etc.).

In practice, our approach involves collecting on-policy data that captures examples of typical errors of the model from the first alignment stage. We then perform post-editing of those on-policy translations, forming a contrastive dataset where post-edits serve as positive examples and the original model outputs as negative ones. We also perform an additional adaptation for WMT topics and construct a dataset derived from WMT24 data annotated with the RATE protocol (Popov et al., 2025). We create contrastive pairs by selecting translations where the average of fluency and accuracy scores differs by more than 5 points, as detailed in Section 2.5.1. The structure preservation dataset, also described in Section 2.5.1, similarly features targeted differences between negative and positive examples, focusing on maintaining specific structural elements rather than completely rewriting translations.

2.5.1 Datasets

Human post-editing

Following active learning paradigm, we collect human-written post-edits of first-stage alignment model generations. This naturally results in triplets of ("source text", "model translation", "human post-edit"). Hereby, the triplets contain a more concentrated signal that specifically targets model error corrections.

On-policy side-by-side comparisons

In addition to human-written post-edits, we collect side-by-side comparisons of on-policy model

generations. The generations are obtained via standard sampling with temperature of 1.0 and side-by-side evaluation is conducted by professional annotators.

The impact of these two portions of the data can be formalized as

(a) eliminating systematic bias of the model in the form of error correction done by human

(b) decreasing variance of the model by training on on-policy comparisons

Structure preserving training

In order to translate structured data, such as HTML web pages, documents with formulae blocks, or subtitles with clear replics borders, we specifically train the model for tagged data translation.

Following Elshin et al. (2024), we convert all tags or separators to the universal tag ({ for the opening paired tag and } for the closing paired tag, each unpaired tag is converted to {})

Consequently, each input containing tags would first be converted to the universal tag format, for example, "<title>Paper Index - EMNLP 2021 Sixth Conference on Machine Translation (WMT21)</title>" would correspond to "{Paper Index - EMNLP 2021 Sixth Conference on Machine Translation (WMT21)}" input during inference.

This implies that for the correct translation of tagged data, the model should be able to preserve the exact bracket sequence given at the input, the number and the sequence of brackets.

We explicitly train for this property using a rule-based reward; for each (source, translation) pair it is algorithmically possible to determine whether the given translation has correctly preserved the tag sequence. Hence, one can construct training samples in an unsupervised way using diverse decoding or beam search:

1. Sample a set of model generations using diverse decoding or wide beam search
2. Score all outputs using rule-based reward that verifies the tag preservation
3. If possible, create a contrastive triplet wherein the positive example contains the output with correctly preserved tags and the negative one contains tag errors

Aiming to keep only informative samples, we select only the sources with tag error in the top-1 model hypothesis and take the most probable example with correctly translated tags to maximize the general translation quality.

WMT24 data

For domain adaptation purposes, we leverage a dataset annotated using the RATE protocol (Popov et al., 2025). This dataset encompasses all documents from WMT24 General Translation Task along with translations from 8 systems participating in the contest, comprising approximately 4,000 segments. The RATE protocol provides detailed information about error spans as well as pointwise accuracy and fluency scores on a 100-point scale.

To generate a contrastive training set from this resource, we take all translation pairs and select only those where the average of fluency and accuracy scores differs by more than 5 points. This selection process yields a dataset containing slightly over 7,000 contrastive examples. It's worth noting that the resulting dataset contains triplets that share the same source text, and, in some cases, a system translation may serve as a positive example in one triplet while appearing as a negative example in another, depending on the quality of alternatives it's being compared against.

2.6 Decoding

We employ a mixed decoding strategy by merging paragraph sequences into larger decoding chunks, hereby decoding with a **local context**.

Inference with Local Context

It is clear that an accurate translation of the document should be done with its full context. However, in practice we have observed that current translation models exhibit inferior performance when given inputs of sufficiently large size.

Given that translation quality starts to decline from several hundred tokens, we propose a hybrid decoding strategy.

1. Set decoding block size to 100 tokens.
2. For a given document, merge sequential paragraphs into blocks greedily, while the currently accumulated block size is less than the decoding block size.
3. Consider the current sequence of paragraphs a single decoding "block".
4. Next blocks are constructed accordingly.

Model	METRICX-24-XXL	METRICX-24-XXL-QE	Fluency
SFT (PTune)	4.272	3.854	0.795
CPO Stage 1	2.417	2.167	0.902
DPO Stage 2	2.369	2.136	0.895
DPO Stage 2 + tags	2.406	2.192	0.898
DPO Stage 2 + tags + WMT	2.253	2.171	0.911

Table 2: Model quality dynamics through alignment stages (WMT24 testset).

	WMT24		WMT25	
domain	segments	documents	segments	documents
literary	63	7	6	2
news	94	16	41	14
social	272	33	27	9
speech	66	66	62	62

Table 3: Testsets descriptive statistics.

To preserve the paragraph structure, we wrap each paragraph in the block with {} tags, thus making sure that the translated document would contain the same number of paragraphs as the source document.

Table 2 shows the quality dynamic of alignment components on WMT24 testset. The first alignment stage displays a significant improvement over SFT with a large margin on all metrics, whereas subsequent alignment stages do not bring statistical improvements in MetricX. We hypothesize that automatic evaluation becomes insensitive from a certain quality of translations and does not capture subtle differences that human annotators still observe.

3 Human Evaluation Results

Due to constraints on time and human resources we selected subsets from two testsets for human evaluation: 495 segments from WMT24 general MT testset and 136 segments from WMT25 general MT blindset. These subsets are hereafter referred to as WMT24 and WMT25, respectively. Obviously, the human translations of WMT25 testset could not be included in this evaluation campaign since they have not been publicly released yet. The number of segments and their distribution across domains is presented in Table 3.

Two protocols were implemented to evaluate MT quality of our submission: the ESA protocol (Kocmi et al., 2024), following WMT guidelines, and the RATE protocol, introduced in Popov et al. (2025). The annotators' qualifications and detailed annotation setup are described in Appendix A.

Yandex+WMT is compared to Yandex model (referred to as DPO Stage 2 + tags above) and several LLM translations. Claude3.7 and GPT-4 translations for WMT24 testset were obtained directly from WMT24 publicly released data. The WMT25 testset was translated using Claude 4 and GPT-4.1 with a simple prompt *"You are a professional English-to-Russian translator. Your goal is to accurately convey the meaning and nuances of the original English text while adhering to Russian grammar, vocabulary, and cultural sensitivities. Produce only the Russian translation, without any additional explanations or commentary. Translate the following text: {input text}"*.

We report segment-level ESA scores alongside error counts and MQM-like scores calculated as $5 \times major + minor$ in Table 4. For the RATE protocol, separate scores for accuracy, fluency, and style are reported, as well as error category statistics. We also report error-per-token statistics and macro-averaged counts by document, domain, or both in Appendix B. Following the WMT methodology, we compute pairwise statistical significance of the differences by Wilcoxon signed rank test and group systems into clusters represented numerically in the tables. Both evaluation methods on WMT25 testset confirm that Yandex+WMT outperforms Yandex and demonstrates statistically significant improvement over the compared LLMs on both testsets. RATE results provide more interpretable differentiation between Yandex and Yandex+WMT: the fluency score shows a measurable increase, while

	segment level					counts per token		
system	errors	major	minor	MQM	ESA	errors	major	minor
Claude 3.7	2.35_3	0.79_2	1.56_3	5.65_5	79.48_3	0.09	0.03	0.06
GPT-4.1	2.64_4	1.10_3	1.54_3	7.34_3	75.41_5	0.11	0.05	0.06
RefA	2.20_2	1.10_3	1.10_2	7.44_4	78.38_4	0.10	0.05	0.05
Yandex	1.51_1	0.74_1	0.76_1	5.06_1	82.16_1	0.07	0.03	0.04
Yandex+WMT	1.50_1	0.80_2	0.70_1	5.35_2	81.30_2	0.07	0.04	0.03

Table 4: Segment-level ESA annotation results on WMT24 testset.

	segment level					counts per token		
system	errors	major	minor	MQM	ESA	errors	major	minor
Claude 4	5.94_3	2.77_3	3.18_4	17.22_3	68.05_4	0.06	0.03	0.03
GPT-4.1	4.67_2	2.56_2	2.11_3	15.69_2	71.77_3	0.05	0.03	0.02
Yandex	4.02_1	2.42_1	1.60_2	15.51_1	72.19_2	0.04	0.03	0.02
Yandex+WMT	3.69_1	2.37_1	1.32_1	15.46_1	73.51_1	0.04	0.02	0.01

Table 5: Segment-level ESA annotation results on WMT25 testset.

differences in accuracy remain statistically non-significant.

4 Conclusion

In this paper, we describe Yandex submission to the WMT25 General Translation task. For English-to-Russian translation direction, our model outperforms all systems from WMT24 competition and achieves results comparable to major foundational LLMs on WMT25 benchmark, as measured by ESA and RATE human evaluation protocols. According to the official human evaluation results our model achieves parity with human reference translations on ESA score.

We present a detailed description of our training procedure as well as giving the rationale for different training steps. Our pipeline includes multi-stage alignment procedure specifically designed to improve the quality of an already capable machine translation system, with the performance close to the human one.

We employ novel techniques, such as curriculum learning with sentences-to-documents adaptation, training on human post-edits and fine-tuning for tagged data using rule-based reward. Our results, measured both in automatic metrics and human evaluations, demonstrate the effectiveness of the proposed pipeline as well as high overall translation quality of the system.

	segment level							
system	errors	major	minor	MQM	accuracy	fluency	style	RATE
Claude 4	13.81_3	2.78_2	10.03_4	23.96_2	70.78_2	58.90_4	89.06_3	23.81_3
GPT-4.1	10.94_2	2.08_1	7.86_3	18.38_1	73.69_1	69.04_3	92.04_2	18.33_2
Yandex	9.07_1	2.21_1	5.68_2	17.77_1	70.42_2	74.39_2	92.94_2	$16.95_1 f$
Yandex+WMT	8.76_1	2.08_1	5.00_1	16.93_1	68.64_3	77.46_1	94.82_1	16.11_1

Table 6: Segment-level RATE annotation results on WMT25 testset.

References

Duarte M. Alves, José P. Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *ArXiv*, abs/2402.17733.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *International Conference on Machine Learning*.

Denis Elshin, Nikolay Karpachev, Boris Gruzdev, Ilya Golovanov, Georgy Ivanov, Alexander Antonov, Nickolay Skachkov, Ekaterina Latypova, Vladimir Layner, Ekaterina Enikeeva, Dmitry Popov, Anton Chekashev, Vladislav Negodin, Vera Frantsuzova, Alexander Chernyshev, and Kirill Denisov. 2024. From general llm to translation: How we dramatically improve translation quality using human evaluation data for llm finetuning. In *Conference on Machine Translation*.

Miquel Esplà-Gomis. 2009. Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In *Beyond Translation Memories: New Tools for Translators Workshop*, Ottawa, Canada.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. pages 492–504.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ArXiv*, abs/2110.07602.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. Gpt understands, too. *AI Open*, 5:208–215.

Dmitry Popov, Vladislav Negodin, Ekaterina Enikeeva, Iana Matrosova, Nikolay Karpachev, and Max Ryabinin. 2025. Refined assessment for translation evaluation: Rethinking machine translation evaluation in the era of human-level systems. Under review.

Brian Thompson. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *ArXiv*, abs/2401.08417.

domain	system	errors	major	minor	MQM	ESA
literary	**Claude 4**	2.39	**0.78**	1.61	**5.50**	82.78
	GPT-4.1	2.72	0.89	1.83	6.56	75.78
	Yandex	**1.67**	1.00	**0.67**	7.89	**87.56**
	Yandex+WMT	2.67	1.72	0.94	10.67	77.67
news	**Claude 4**	4.15	2.04	2.11	12.35	72.76
	GPT-4.1	3.50	1.83	1.67	11.19	74.67
	Yandex	3.34	1.96	1.38	12.24	73.49
	Yandex+WMT	**2.38**	**1.58**	**0.80**	**10.01**	**79.65**
social	**Claude 4**	5.93	2.47	3.46	15.86	68.00
	GPT-4.1	4.25	2.06	2.19	13.11	73.51
	Yandex	**3.38**	**2.02**	**1.36**	**12.17**	**73.95**
	Yandex+WMT	3.80	2.42	1.38	15.53	71.77
speech	**Claude 4**	7.48	3.57	3.91	22.17	63.53
	GPT-4.1	5.81	3.42	2.39	20.68	68.72
	Yandex	4.97	3.03	1.95	19.87	69.07
	Yandex+WMT	**4.61**	**2.94**	**1.68**	**19.49**	**69.82**

Table 7: WMT25 ESA annotation results - segment-level average by domains.

A Annotation details

The evaluation was conducted by two distinct groups of annotators:

1. ESA experts: an in-house group of Russian language natives who successfully passed the C1-level English test and regularly participate in translation evaluation campaigns. These experts received training to annotate translations according to ESA instructions. Quality control was maintained through manually prepared golden annotations.

2. RATE experts: a smaller in-house group of Russian language natives with qualification in Linguistics or Translation who underwent a multi-step selection process based on translation, post-editing and fact-checking competencies. As shown in Popov et al. (2025), this rigorous selection procedure may be crucial for ensuring annotation quality when evaluating high-quality translations.

B Extended human evaluation results

B.1 WMT25 ESA results

system	errors	major	minor	MQM	ESA
Claude 4	4.99	2.21	2.77	13.97	71.77
GPT-4.1	4.07	**2.05**	2.02	12.88	73.17
Yandex	**3.34**	**2.00**	1.34	**13.04**	**76.02**
Yandex+WMT	**3.37**	**2.16**	**1.20**	**13.93**	74.72

Table 8: WMT25 ESA annotation results - segment-level values macro averaged by domains.

system	errors	major	minor	MQM	ESA
Claude 4	6.66	3.15	3.52	19.55	65.90
GPT-4.1	5.20	2.96	2.24	18.03	70.34
Yandex	4.47	2.70	1.77	17.56	70.69
Yandex+WMT	**4.12**	**2.63**	**1.49**	**17.33**	**71.81**

Table 9: WMT25 ESA annotation results - document-level scores.

system	errors	major	minor	MQM	ESA
Claude 4	4.98	2.22	2.76	13.97	71.73
GPT-4.1	4.06	2.05	2.02	12.86	73.17
Yandex	**3.34**	**2.00**	1.34	**13.03**	**75.99**
Yandex+WMT	**3.36**	2.16	**1.21**	13.89	74.77

Table 10: WMT25 ESA annotation results - document-level scores macro-averaged by domain.

domain	system	errors	major	minor	MQM	accuracy	fluency	style	RATE
literary	**Claude 4**	8.83	1.42	6.33	13.42	80.08	68.58	94.17	13.33
	GPT-4.1	8.83	**0.67**	7.17	10.50	86.42	74.33	**95.33**	11.50
	Yandex	**5.08**	0.92	**3.75**	**8.33**	80.08	**82.00**	94.58	**8.75**
	Yandex+WMT	6.58	0.83	4.25	8.83	**83.58**	80.92	94.50	9.50
news	**Claude 4**	10.41	2.09	7.52	18.01	70.09	67.20	88.57	17.72
	GPT-4.1	7.91	1.61	5.48	13.65	**76.39**	75.10	93.66	13.61
	Yandex	**6.37**	1.72	**3.77**	12.91	73.83	79.57	93.29	11.83
	Yandex+WMT	6.84	**1.39**	3.83	**12.55**	72.33	**82.01**	**95.35**	**11.96**
social	**Claude 4**	11.78	2.17	9.11	19.94	73.44	63.09	89.78	19.85
	GPT-4.1	10.26	1.61	7.70	15.85	**75.78**	71.46	91.39	16.20
	Yandex	8.91	**2.04**	5.67	16.22	71.94	75.00	91.37	15.44
	Yandex+WMT	**7.83**	2.09	**4.46**	**15.76**	68.41	**79.39**	**94.59**	**15.30**
speech	**Claude 4**	17.42	3.63	12.44	30.67	69.17	50.65	88.58	30.57
	GPT-4.1	13.44	2.73	9.57	23.36	**69.77**	63.48	90.94	23.04
	Yandex	11.31	2.74	7.13	22.57	66.56	69.97	93.23	21.78
	Yandex+WMT	**10.65**	**2.66**	**6.07**	**21.11**	64.85	**73.27**	**94.60**	**19.85**

Table 11: WMT25 RATE annotation results - segment-level scores by domain.

system	errors	major	minor	MQM	accuracy	fluency	style	RATE
Claude 4	12.11	2.32	8.85	20.51	**73.20**	62.38	90.27	20.37
GPT-4.1	10.11	1.65	7.48	15.84	77.09	71.09	92.83	16.09
Yandex	**7.92**	1.85	5.08	15.01	73.11	76.64	93.12	14.45
Yandex+WMT	7.98	**1.74**	**4.65**	**14.56**	72.29	**78.90**	**94.76**	**14.15**

Table 12: WMT25 RATE annotation results - segment-level scores macro-averaged by domain.

B.2 WMT25 RATE results

system	errors	major	minor	MQM	accuracy	fluency	style	RATE
Claude 4	15.50	3.17	11.16	27.09	70.02	55.00	88.85	26.97
GPT-4.1	12.11	2.38	8.65	20.70	**71.88**	66.47	91.53	20.52
Yandex	10.13	2.46	6.36	20.04	68.58	72.29	93.07	19.23
Yandex+WMT	**9.64**	**2.35**	**5.50**	**18.87**	66.89	**75.45**	**94.73**	**17.85**

Table 13: WMT25 RATE annotation results - document-level scores.

system	errors	major	minor	MQM	accuracy	fluency	style	RATE
Claude 4	12.10	2.32	8.84	20.46	73.21	62.36	90.30	20.33
GPT-4.1	10.09	1.65	7.46	15.80	**77.16**	71.17	92.83	16.05
Yandex	**7.92**	1.86	5.08	15.02	73.08	76.61	93.10	14.46
Yandex+WMT	**7.96**	**1.74**	**4.64**	**14.52**	72.34	**78.8**4	**94.79**	**14.12**

Table 14: WMT25 RATE annotation results - document-level scores macro-averaged by domain.

severity	category	Claude 4	GPT-4.1	Yandex	Yandex+WMT
major	Consistency	0.11	0.09	0.16	0.17
	Do not translate	0.03	0.02	0.03	0.03
	Fluency	1.09	0.57	0.43	0.31
	Grammar	0.03	0.03	0.00	0.01
	Mistranslation	1.37	1.21	1.38	1.42
	NE	0.14	0.08	0.10	0.12
	Omission	0.00	0.00	0.00	0.00
	Overtranslation	0.01	0.06	0.05	0.10
	Punctuation	0.00	0.01	0.00	0.00
	Style	0.01	0.01	0.01	0.00
	Undertranslation	0.08	0.07	0.10	0.08
minor	Do not translate	0.06	0.02	0.01	0.01
	Fluency	6.19	4.85	3.59	2.80
	Grammar	1.26	0.64	0.22	0.27
	Mistranslation	1.39	1.33	1.19	1.03
	NE	0.17	0.17	0.11	0.18
	Omission	0.00	0.00	0.00	0.00
	Overtranslation	0.02	0.10	0.05	0.15
	Punctuation	0.31	0.14	0.02	0.04
	Style	0.33	0.33	0.18	0.20
	Undertranslation	0.16	0.21	0.24	0.20
trivial	Do not translate	0.00	0.00	0.00	0.00
	Fluency	0.63	0.61	0.43	0.53
	Grammar	0.02	0.03	0.02	0.05
	Mistranslation	0.06	0.05	0.04	0.07
	NE	0.00	0.01	0.04	0.02
	Omission	0.00	0.00	0.00	0.00
	Overtranslation	0.01	0.01	0.00	0.01
	Punctuation	0.01	0.01	0.00	0.00
	Style	0.02	0.03	0.02	0.01
	Undertranslation	0.01	0.00	0.02	0.01

Table 15: WMT25 RATE annotation results - segment-level error counts grouped by severity (major - 4-5, minor - 2-3, trivial - 1).

IRB-MT at WMT25 Translation Task: A Simple Agentic System Using an Off-the-Shelf LLM

Ivan Grubišić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
name.surname@irb.hr

Damir Korenčić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
name.surname@irb.hr

Abstract

Large Language Models (LLMs) have been demonstrated to achieve state-of-the-art results on machine translation. LLM-based translation systems usually rely on model adaptation and fine-tuning, requiring datasets and compute. The goal of our team's participation in the "General Machine Translation" and "Multilingual" tasks of WMT25 was to evaluate the translation effectiveness of a resource-efficient solution consisting of a smaller off-the-shelf LLM coupled with a self-refine agentic workflow. Our approach requires a high-quality multilingual LLM capable of instruction following. We select `Gemma3-12B` among several candidates using the pretrained translation metric `MetricX-24-XL` and a small development dataset. WMT25 automatic evaluations place our solution in the mid tier of all WMT25 systems, and also demonstrate that it can perform competitively for approximately 16% of language pairs.

1 Introduction

Machine translation (MT) is an important yet unsolved NLP task with significant practical applications Kocmi et al. (2024a). Large language models (LLMs) have become a basis for state-of-the-art MT solutions, and the best approaches rely either on commercial LLMs or on open-weights models adapted using translation-specific data (Kocmi et al., 2024a). However, commercial cloud-based models may introduce cost constraints and dependency issues, while the adaptation of open-weight models commonly requires substantial computational resources and specialized training datasets.

Recent developments in LLMs yielded smaller yet capable models such as Gemma3 (Team et al., 2025) and Qwen3 (Yang et al., 2025), which are multilingual and support instruction following and reasoning. In parallel, research in multi-agent systems led to task-independent workflows, such as self-refine (Madaan et al., 2023), and task-oriented workflows where individual agents assume natural task-specific roles (Wu et al., 2024). Both approaches have demonstrated the capability of the agentic workflows to outperform individual LLMs (Madaan et al., 2023; Wu et al., 2024).

We hypothesize that the combination of capable smaller LLMs and agentic workflows has the potential to create a resource-effective translator with solid performance. Our participation in the WMT25 Translation Task (Kocmi et al., 2025a) is oriented toward testing this hypothesis in the controlled environment of the "constrained" track, which allows only openly available datasets and models below 20B parameters.

We evaluate our approach on the WMT25 General Machine Translation task, which assesses MT systems across four domains (news, social media, speech, and literary) with document-level context and multi-modal resources including video, image, and speech data (Kocmi et al., 2025a). The task comprises 16 language pairs covering major language groups including morphologically rich, low-resource, and diverse script languages. We also participate in the Multilingual subtask, which extends evaluation to 15 additional target languages. Overview of the dataset statistics can be found in Table 3.

As the first step in designing our system we tested several multilingual generative models and encoder-decoder models specialized for translation, evaluating them on a subset of pairs from the WMT24++ dataset (Deutsch et al., 2025). `Gemma3-12B` (Team et al., 2025) proved to be the best solution in terms of `MetricX-24-XL` metric (Juraska et al., 2024) so our final system is based on this model. We enhance the model with a version of the self-refine workflow (Madaan et al., 2023) based on a prompt adapted for machine translation. Our system uses as input only the text modality, and works with paragraph-sized text segments.

*Equal contribution.

Proceedings of the Tenth Conference on Machine Translation, pages 753–764
November 8-9, 2025

WMT25 evaluations using a number of automatic translation metrics (Kocmi et al., 2025b) show that our system achieves mid-level performance when compared with all the participating systems that include team-submitted solutions (both constrained and unconstrained), as well as benchmarks added by the organizers (individual LLMs and commercial solutions). The system achieves competitive performance for five language pairs (en→zh, en→de, en→id, en→sv, en→vi) and human ESA annotations (Kocmi et al., 2025a) show that it often generates good translations. We make the code of the system freely available.[1]

2 Related Work

Several multi-agent LLM systems for machine translation (MT) have been proposed recently, often inspired by human collaborative problem-solving and professional translation workflows (Wu et al., 2024; Peter et al., 2024; Briakou et al., 2024; Wang et al., 2025b; Anonymous, 2025). These systems aim to address the limitations of single-model MT systems, including in handling linguistic nuances, context, and idiomatic expressions. They consist of autonomous LLM-based agents assigned to specialized tasks, organized in a workflow and sometimes embedded in an iterative loop.

Such multi-agent systems can demonstrate superior performance compared to non-agentic baselines (Briakou et al., 2024; Wang et al., 2025a; Anonymous, 2025), . For example (Briakou et al., 2024) reported large improvements over conventional zero-shot prompting and even outperformed top-performing WMT 2024 systems in some cases. Furthermore, human evaluations frequently show a preference for translations produced by these multi-agent systems (Wu et al., 2024; Anonymous, 2025).

While effective, the proposed systems rely on powerful LLMs (such as GPT-4o and Gemini 1.5 Pro) and often involve complex and computation-intensive workflows, which entail latency and computational overhead. In contrast, our system relies on a smaller open-weights model and a simple one-step self-refine workflow (Madaan et al., 2023). This makes it resource efficient with translation time comparable to zero-shot inference with a single LLM.

[1] https://github.com/igrubi/irb-mt-wmt2025

3 Dataset

The WMT25 translation evaluation dataset encompasses 31 language pairs with diverse characteristics, enabling assessment across different language families, resource levels, and translation scenarios (Kocmi et al., 2025a). The list of language pairs and the statistics of associated sub-datasets is displayed in Table 3.

The General MT task comprises 16 language pairs covering both large and small languages. The task includes both English-centric and non-English language pairs, with English-to-target directions covering Arabic (Egyptian), Bhojpuri, Chinese (Simplified), Czech, Estonian, Icelandic, Italian, Japanese, Korean, Maasai (Kenya), Russian, Serbian (Latin), and Ukrainian. Additionally, the task features non-English source languages with Czech-to-German, Czech-to-Ukrainian, and Japanese-to-Chinese pairs.

The dataset exhibits significant variation in size and text complexity. Russian dominates with 7,804 texts, followed by Hindi with 5,087 texts, while smaller language pairs like Italian contain only 87 texts. The dataset contains texts from four domains: news articles, transcripts of video speech associated with audio data, social media posts associated with printscreen images, and literary texts. A significant number of texts from the General MT subtask does not belong to any domain.

The Multilingual (sub)task extends evaluation to 15 additional target languages: Bengali, German, Greek, Persian, Hindi, Indonesian, Italian, Kannada, Lithuanian, Marathi, Romanian, Serbian (Cyrillic), Swedish, Thai, Turkish, and Vietnamese. English is the only source language, and all the texts belong to one of the four domains described above.

As the statistics in Table 3 show, the General MT texts tend to be short, predominantly with 100–200 tokens, and mostly consist of a single paragraph. The Multilingual subtask texts are longer (with the exception of Hindi), having over 400 tokens on average, and tend to consist of at least several paragraphs that are longer than the General MT paragraphs. Pair with the largest document collection (English-Hindi) is an exception since it contains mostly short texts.

4 System

The goal of our submission was to examine how a lightweight, simple, and computationally effi-

Qwen3		NLLB		Gemma3		EuroLLM		Aya	
Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
6.98	3.40	6.54	3.70	4.59	2.59	5.03	2.83	5.01	2.88

Table 1: Average performance of LLMs benchmarked on the development set, measured by `MetricX-24-XL` (lower is better).

cient agentic workflow for MT compares against a range of other approaches . To enable a fair comparison with systems utilizing a similar level of resources, we submitted our solution to the "constrained" track (Kocmi et al., 2025a) which allows only open-weights models with a combined size below 20B parameters.

We always apply our translation methods on the level of a paragraph, i.e., no document-wide context is used. Additionally, only text is used as input, i.e., image and speech data provided for parts of the dataset is not used.

4.1 Model Selection

A self-refine (Madaan et al., 2023) agentic workflow for MT requires a translation model and a model able to implement the refinement of a translation. For both steps strong multilingual capabilities are desirable, and the refinement model should have instruction following capabilities in order to properly implement the refinement instructions.

In order to select the appropriate models we tested several LLMs on a subset of language pairs from the WMT24++ dataset (Deutsch et al., 2025), which we used to create the development dataset. To this end, we used pairs from WMT24++ that matched pairs from the WMT25 General MT track (see Table 3 for a list of WMT25 pairs). Since WMT24++ contains only en→X pairs, we created the non-English development pairs (cs→uk, cs→de, ja→zh) by matching via English texts. WMT24++ does not contain English-Bhojpuri (en→bho) and English-Massai (en→mas) data, so we did not test on these pairs. The final development set was constructed by subsampling 200 texts for each language pair.

We tested the following models: `Gemma3-12B` , `EuroLLM-9B-Instr` , `Qwen3-8B` , `Aya-101` , and `NLLB-200-3.3B` (Team et al., 2025; Martins et al., 2025; Yang et al., 2025; Üstün et al., 2024; NLLB Team, 2024). `Gemma3-12B` , `Qwen3-8B` and `EuroLLM-9B-Instr` are modern implementations of the GPT architecture, supporting multilinguality, instruction following, and, in the case of `Gemma3-12B` and `Qwen3-8B` , reasoning. These models can serve as basis for both the translator and the translation refinement agent. `Aya-101` is a massively multilingual instruction following model, but in this context we view it purely as a translation model since its effective input context length of 1024 limits the applicability to expectedly longer refinement prompts. `NLLB-200-3.3B` is a state-of-the-art encoder-decoder transformer trained specifically for multilingual machine translation. All models except `NLLB-200-3.3B` (trained to translate the entire input text) were equipped with simple translation prompts detailed in Appendix C.1.

`MetricX-24-XL` (Juraska et al., 2024) (the "metricx-24-hybrid-xl-v2p6" variant) was used to estimate the models' performance. `MetricX-24-XL` MetricX is a metric learned from parallel text with source, hypothesis, and reference segments that are annotated with human Direct Assessment (DA) and MQM scores. It is based on the mT5 transformer model (Xue et al., 2021) with a regression head, and it can estimate translation quality both with and without a reference translation.

Table 1 shows that Gemma3 has the best average translation performance (averaged over all language pairs), and that it has lowest average standard deviation among all tested models. This indicates that is should have the best and most stable translation performance. Per-pair results in Table 4 in Appendix B show that Gemma3 has superior or competitive performance for almost all language pairs. Additionally, Gemma3 has both instruction-following and reasoning capabilities (Team et al., 2025).

For these reasons, we decided that it is an optimal model for both translation and refinement. Additional benefit of this choice is the use of a single model for the entire workflow, which reduces the memory footprint.

4.2 The Agentic Workflow

The final workflow implements a two-stage translation process based on the self-refine workflow

Table 2: AutoRank translation scores formed by aggregating multiple automatic translation metrics (Kocmi et al., 2025b) and data on the relative position of IRB-MT, for both GeneralMT (top row group) and Multilingual (bottom row group) tracks. For each pair, AutoRank scores for IRB-MT and `Gemma3-12B` are given. Additionally, for each system subcategory, the total number of systems (#sys) and the number of systems ranked above IRB-MT (#above) is given. Constrained systems use openly available data and models below 20B parameters. Team systems are submitted by participating teams, while Benchmark systems are included by the organizers (Kocmi et al., 2025b).

Lang.	**IRB-MT**	**Gemma3**	**Constrained**				**Unconstrained**			
Pair			**Team**		**Benchmark**		**Team**		**Benchmark**	
			# sys	**# above**	**# sys**	**# above**	**# sys**	**# above**	**# sys**	**# above**
cs→de	12.1	**11.2**	9	5	9	2	7	3	15	13
cs→uk	**8.9**	9.7	10	6	9	1	8	3	15	11
ja→zh	**12.1**	17.1	9	6	9	1	8	6	15	9
en→ar	**10.8**	11.7	7	5	9	1	6	3	15	12
en→bho	**11.4**	12.3	7	4	9	1	6	3	13	10
en→zh	**9.3**	10.6	9	5	9	0	5	3	15	9
en→cs	**12.6**	13.4	12	8	9	1	6	3	15	13
en→et	**11.1**	12.1	8	7	9	0	6	4	15	9
en→is	**11.9**	13.8	4	3	9	1	6	5	14	9
en→it	**10.2**	15.5	4	2	9	0	5	3	15	11
en→ja	**10.3**	13.6	11	8	9	0	8	5	15	12
en→ko	**8.4**	9.0	7	4	9	0	5	3	15	8
en→mas	9.7	**8.8**	2	1	9	7	5	3	11	9
en→ru	**9.9**	14.7	9	6	9	0	8	4	14	6
en→sr	**6.3**	7.6	7	5	9	0	6	2	13	5
en→uk	**8.0**	14.4	9	5	9	0	9	3	15	6
en→bn	**5.1**	7.6	2	1	9	0	5	2	14	5
en→de	**9.8**	12.2	3	1	9	1	5	4	15	13
en→el	**5.9**	9.9	3	2	9	0	5	3	15	8
en→fa	**5.1**	5.7	2	1	9	0	5	3	14	8
en→hi	**5.3**	7.1	2	1	9	0	5	3	14	5
en→id	**5.5**	6.6	2	1	9	0	5	3	15	6
en→kn	**11.0**	13.4	2	1	9	0	5	4	14	9
en→lt	**8.9**	10.2	3	2	9	0	5	4	15	9
en→mr	**7.3**	12.4	2	1	9	0	5	3	14	6
en→ro	**6.4**	7.9	3	1	9	1	5	3	15	8
en→sr_Cy	**9.9**	12.1	4	2	9	0	6	5	13	5
en→sv	**5.8**	11.4	3	1	9	0	5	3	15	6
en→th	**4.8**	9.1	2	1	9	0	5	2	14	7
en→tr	**7.2**	8.7	2	1	9	0	5	3	15	7
en→vi	**5.1**	8.1	2	1	9	0	5	3	14	6

(Madaan et al., 2023). The initial translation is generated using the provided WMT25 prompts (slightly modified by dropping the instruction to respect the paragraphs structure). Details of the prompts used for the translation workflow can be found in Appendix C.2.

In the next step a refinement prompt, tailored for machine translation, is executed. The refinement prompt consists of the original translation prompt, the input text, the initial translation, and task-specific instructions. The instructions elicit the model to reason about the improvement and to produce the solution enclosed within "`<solution>` `</solution>`" tags. For efficiency, the model is instructed to keep the reasoning at "close to 300 words". The inference temperature was set to 0 (no sampling), and the maximum number of new tokens was set to 20K.

We evaluated the self-refine workflow by comparing it with the basic `Gemma3-12B` translator on both General MT and Multilingual language pairs from the WMT25 dataset. `MetricX-24-XL` scores showed that the self-refine approach has similar or slightly better scores across the majority of language pairs. We took this as evidence that the proposed agentic system does not perform worse then the baseline. Since the original self-refine experiments show improvements for a number of models and tasks (Madaan et al., 2023), we were confident that the agentic system would best the base translator when evaluated with other translation metrics.

5 Results

IRB-MT submitted translations for all of the 31 language pairs of the GeneralMT and Multilingual subtasks (Kocmi et al., 2025a). Automatic non-human evaluations show that IRB-MT is a mid-tier constrained system that outperforms baseline `Gemma3-12B` in most cases, which demonstrates the benefit of the self-refine approach. Out of the 16 pairs for which human evaluation was performed, IRB-MT's performance was high enough for it to be selected for human evaluation in the case of 13 pairs (81.25% of pairs (Kocmi et al., 2025b)).

To further analyze the relative performance of IRB-MT we rely on the AutoRank metric that aggregates the rankings of multiple translation quality metrics, in order to mitigate biases of individual metrics (Kocmi et al., 2025b). Other evaluated translation systems consist of systems submitted by participating teams, and additional organizer-chosen systems included for comparison (Kocmi et al., 2025b). We label these two groups "Team" and "Benchmark" systems, respectively. Benchmark systems include open-weight and cloud-based LLMs, and commercial MT systems.

The AutoRank statistics in Table 2 show that IRB-MT outperforms Gemma3 for all but two language pairs. In the constrained track, IRB-MT compares favorably with the Benchmark systems, being above most of them. This is not surprising since these are smaller (below 20B parameters) LLMs applied as zero-shot translators. On the other hand, when compared to Team constrained systems IRB-MT is, for most pairs, located at or below the median rank. Presumably, these systems are mostly data-based, i.e., they rely on model adaptation and fine-tuning.

As for the unconstrained track IRB-MT compares relatively favorably with the Team systems, often placed close to the middle of the list. However, this probably has most to do with the fact that, surprisingly, submitted unconstrained systems generally perform worse than the submitted constrained systems (Kocmi et al., 2025b). Unconstrained Benchmark systems consist of mid-sized LLMs (above 20B parameters), commercial LLMs, and commercial translation systems. These systems mostly outperform IRB-MT for large languages and for most European languages, while IRB-MT compares more favorably and sometimes competitively on mid-and lower-resourced languages and non-European languages.

As the table Table 2 contains only information on relative performance, we provide statistics on the GEMBA-ESA translation scores (Kocmi and Federmann, 2023) computed using GPT4.1 (Kocmi et al., 2025b). The scores, contained in Table 5 in Appendix D, lay out the scores of IRB-MT and top-performing systems from all categories. IRB-MT scores range between approx. 50 and approx. 75 for most systems (100 being the perfect performance). However, for 5 language pairs (en→zh, en→de, en→id, en→sv, en→vi) IRB-MT both achieves a score close to 80 and is approximately 10 points below the top-performing system, which shows the potential of the approach.

Human evaluation (Kocmi et al., 2025a) of selected translation systems was performed by applying the ESA annotation method (Kocmi et al., 2024b) and, for two language pairs, by applying the MQM method (Freitag et al., 2021). The evaluated systems were clustered based on the statistical significance of performance differences (Kocmi et al., 2025a).

When compared to other human-evaluated systems, IRB-MT is located at or below the median for the majority of language pairs. Out of 11 pairs for which the systems were ESA-annotated, for 6 pairs IRB-MT either has a score above 66%, or it is not significantly different from systems scoring above 66% (Kocmi et al., 2025a). According to ESA annotation guidelines the score of 66% is the threshold for translations with "Most meaning preserved and few grammar mistakes" (Kocmi et al., 2024b). We take this as an argument for our system's ability to generate good translations for a non-trivial percentage of language pairs. For the challenging English-Arabic pair, IRB-MT outperforms all of the constrained systems and compares favorably to 50% of the unconstrained systems (Kocmi et al., 2024b).

6 Conclusion and Future Work

We proposed a simple lightweight "self-refine" workflow for machine translation, based on the multilingual `Gemma3-12B` LLM with instruction-following and reasoning capabilities. Our approach was included in the WMT25 (Kocmi et al., 2025a) evaluation with automatic metrics, performed on a large set of translation systems. The results place our system, on average, in the mid-tier, but there is a significant performance variations across language pairs, with the tendency of better per-

formance on mid- and low-resource languages. While the IRB-MT fails to come close to the top-performing systems, human ESA annotations show that it often produces good translations (Kocmi et al., 2025a).

Future research on the improvement of our approach should tackle the issue of performance variability. Human or LLM-assisted examination of language pairs with the lowest scores would reveal the type of errors and suggest improvements. Although the agentic workflow by-and-large outperforms the base Gemma3 model, there is a significant variation in the gap between the two. Analyzing the language pairs and texts for which IRB-MT fails to improve upon the base system could lead to the refinement of the agentic system.

Other directions for further improvement of IRB-MT include: a larger thinking budget, an iterative improvement loop, and a more granular agentic system with specialized roles. Refinement of the agentic structure could combine elements of the existing MT workflows (Wu et al., 2024; Peter et al., 2024; Briakou et al., 2024; Wang et al., 2025b; Anonymous, 2025). The key challenge is to boost performance without relying on overly complex and long workflows, or on too large LLMs. Examining different combinations of the translator LLM and the refiner LLM could also lead to a performance boost, and to insights into the models' behavior.

In general, it would be interesting to examine how close can purely agentic approaches come to the data-driven approaches based on LLM adaptation and fine-tuning, and how does the compute-vs-performance tradeoff look like.

Acknowledgements

This paper was supported by the European Union's NextGenerationEU program. We would like to thank Tomislav Šmuc, Ph.D., and Prof. Sonja Grgić, Ph.D., for support and valuable discussions. We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 hosted by BSC, Spain, under the project ID EHPC-DEV-2025D05-087.

References

Anonymous. 2025. Agentdiscotrans: Agentic LLMs for discouse-level machine translation. In *Submitted to ACL Rolling Review - February 2025*. Under review.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo

Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of wmt25 general machine translation systems.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–850.

Anishka Peter, Mai Dang, Michael Liu, Joaquin Dominguez, and Nibhrat Lohia. 2024. Multi-agent translation team (matt): Enhancing low-resource language translation through multi-agent workflow. SMU Data Science Review, Vol. 8, No. 3, Article 3. Available at SMU Scholar.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

George Wang, Jiaqian Hu, and Safinah Ali. 2025a. Maats: A multi-agent automated translation system based on mqm evaluation.

Xi Wang, Jiaqian Hu, and Safinah Ali. 2025b. Maats: A multi-agent automated translation system based on mqm evaluation.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024.

TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 15894–15939.

A Dataset Statistics

Table 3: WMT25 Dataset Statistics by Language Pair, for the General MT task (top half), and the "multilingual" subtask (bottom half). For each pair the statistics pertain to the texts in the source language (predominantly English). The statistics include average number of tokens in the text, statistics on the number of paragraphs per text, and on the number of tokens per paragraph. Tokenization is done by splitting on whitespaces, while the paragraphs are separated by a double newline character.

Language Pair	# Texts	Avg Text-Tokens	# Paragraphs			# Para-Tokens		
			Q1	Avg	Q3	Q1	Avg	Q3
cs→uk	230	157.1	1.0	1.8	2.0	64.0	89.2	113.0
cs→de	256	171.1	1.0	1.8	2.0	66.0	95.8	117.0
ja→zh	106	413.8	1.0	3.2	4.0	51.0	131.3	188.0
en→ar	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→bho	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→zh	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→cs	1,277	101.6	1.0	1.2	1.0	39.0	83.8	113.0
en→et	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→is	1,607	84.0	1.0	1.2	1.0	22.0	72.9	101.0
en→it	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→ja	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→ko	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→mas	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→ru	7,804	52.6	1.0	1.0	1.0	13.0	51.0	46.0
en→sr	2,251	137.9	1.0	1.1	1.0	61.0	124.3	184.0
en→uk	1,251	96.3	1.0	1.2	1.0	37.0	80.5	110.0
en→bn	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→de	113	407.7	1.0	3.4	2.0	84.0	120.0	146.0
en→el	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→fa	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→hi	5,087	36.2	1.0	1.0	1.0	18.0	34.6	44.0
en→id	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→kn	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→lt	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→mr	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→ro	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→sr_Cyrl	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→sv	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→th	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→tr	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0
en→vi	87	422.6	1.0	3.8	4.0	82.0	110.8	137.0

B Development set results

Table 4: `MetricX-24-XL` results of benchmarked LLMs on the development set for the language pairs from WMT24++ that occur in the WMT25 General MT track. The models are: `Qwen3-8B` , `NLLB-200-3.3B` , `Gemma3-12B` , `EuroLLM-9B-Instr` , and `Aya-101` .

Language Pair	Qwen3		NLLB		Gemma3		EuroLLM		Aya	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Average	6.98	3.40	6.54	3.70	**4.59**	**2.59**	5.03	2.83	5.01	2.88
cs→de	3.31	2.25	3.45	2.52	2.68	1.97	2.51	1.87	2.73	1.78
cs→uk	7.06	3.39	5.84	3.45	4.40	2.47	4.84	2.67	5.10	2.62
ja→zh	4.44	1.85	5.62	2.96	3.24	1.64	3.40	1.60	3.89	1.96
en→ar	5.46	2.89	5.77	3.49	5.40	2.71	4.82	2.64	5.22	2.81
en→zh	4.18	1.68	5.60	3.50	2.62	1.71	2.88	1.88	3.70	2.25
en→cs	7.87	4.39	7.08	4.36	5.47	3.22	4.85	2.86	6.15	3.51
en→et	15.04	6.10	7.84	4.48	8.08	4.18	6.31	3.62	7.28	4.19
en→is	17.76	6.52	8.79	4.82	10.12	4.75	16.08	7.03	7.42	3.46
en→it	3.43	2.48	3.43	2.60	2.83	2.08	2.80	1.98	4.15	3.13
en→ja	4.55	2.07	7.09	3.34	4.20	2.08	4.46	2.11	4.72	2.18
en→ko	4.82	2.44	14.06	4.11	3.96	1.98	4.32	2.12	4.95	2.79
en→ru	5.05	3.35	5.94	4.54	3.30	2.45	4.50	3.49	4.70	3.13
en→sr	7.60	4.28	4.59	3.42	4.07	2.68	3.85	2.80	4.69	3.15
en→uk	7.14	3.98	6.43	4.22	3.94	2.37	4.74	2.94	5.43	3.29

C Prompts

C.1 Simple Translation Prompts for Model Selection

These prompts were applied for benchmarking `Gemma3-12B` , `EuroLLM-9B-Instr` , `Qwen3-8B` , and `Aya-101` on the development set. In the case of `Aya-101` the system prompt was not used. Only in the case of `Gemma3-12B` was "Output ONLY the translated text!" added to the end of the system prompt, since the initial tests revealed that the model almost always produces "thinking" tokens before producing the translation.

C.1.1 System Prompt

```
You are a professional translator with expertise in multiple languages.
Provide accurate, natural translations that preserve meaning and context.
[Output ONLY the translated text!]
```

C.1.2 User Prompt

```
Please translate the following text from {source_language} to {target_language}:

{text}
```

C.2 Prompts for the Self-Refine Translation Workflow

These prompts were used in combination with `Gemma3-12B` to produce the final translations.

Prompt for `Gemma3-12B` Translator

For the translation agent prompts provided as part of the WMT25 datasets were used, as they are well-formed and convey additional domain-specific information. One such prompt was given for every text in the test dataset. To illustrate the structure of these prompts we display two prompts for Czech-German translation, for the "news" and "social" domains, respectively. The only modification of

the original prompts was the removal of the sentence “Retain the paragraph breaks (double new lines) from the input text.”, which was done because we always applied our translator on individual paragraphs.

You are a professional Czech-to-German translator, tasked with providing translations suitable for use in Germany (de_DE). Your goal is to accurately convey the meaning and nuances of the original Czech text while adhering to German grammar, vocabulary, and cultural sensitivities. The original Czech text is a news article. Ensure the translation is formal, objective, and clear. Maintain a neutral and informative tone consistent with journalistic standards. Produce only the German translation, without any additional explanations or commentary. ~~Retain the paragraph breaks (double new lines) from the input text.~~ Please translate the following Czech text into German (de_DE):
{**text**}

You are a professional Czech-to-German translator, tasked with providing translations suitable for use in Germany (de_DE). Your goal is to accurately convey the meaning and nuances of the original Czech text while adhering to German grammar, vocabulary, and cultural sensitivities. The original Czech text is user-generated content from a social media platform. Ensure you do not reproduce spelling mistakes, abbreviations or marks of expressivity. Platform-specific elements such as hashtags or userids should be translated as-is. Produce only the German translation, without any additional explanations or commentary. ~~Retain the paragraph breaks (double new lines) from the input text.~~ Please translate the following Czech text into German (de_DE):
{**text**}

Prompt for `Gemma3-12B` Translation Refinement

The `reasoning_words` parameter was fixed to 300, and the text of the solution was extracted from the <solution> </solution> tags.

Your job is to review a translation, and correct it if necessary.
You will be given an original text, translation instructions,
and the translation created according to these instructions.
Respect the instructions!

These are the instructions according to which the translation was produced:
{**translation_prompt**}

Original text:
{**original_text**}

Translation:
{**translation**}

First, analyze the instructions, the original text, and the translation.
Then reason about the improved solution (if any), and produce the solution.
Try to keep the reasoning succinct, close to {**reasoning_words**} words!
End with your final solution, enclosed within the <solution> </solution> tags.

D Comparison with other Participating Systems

Table 5: GEMBA-ESA translation scores (Kocmi and Federmann, 2023) computed using GPT4.1 (Kocmi et al., 2025b), for both GeneralMT (top row group) and Multilingual (bottom row group) tracks. Pairs for which GEMBA-ESA were not computed are omitted. For each pair, scores for IRB-MT and `Gemma3-12B` are given, as well as scores for best-performing systems in each sub-category defined by the properties of the system. Constrained systems use openly available data and models below 20B parameters. Team systems are submitted by participating teams, while Benchmark systems are included by the organizers (Kocmi et al., 2025b).

Language	IRB-MT	Gemma3	Constrained		Unconstrained	
Pair			Team	Benchmark	Team	Benchmark
cs→de	75.4	77.5	88.3	77.5	87.5	91.0
cs→uk	74.8	75.9	85.3	76.7	84.3	89.5
ja→zh	70.4	64.1	85.5	69.8	81.8	84.8
en→ar	67.5	67.6	75.0	67.6	75.4	84.5
en→zh	77.5	76.6	88.3	76.6	81.5	88.7
en→cs	73.6	74.1	89.4	75.8	86.2	91.5
en→et	60.5	59.4	87.8	59.4	74.3	90.7
en→is	47.2	42.1	83.9	76.3	85.1	87.6
en→it	79.8	74.7	88.7	78.6	88.0	90.5
en→ja	77.9	73.8	89.6	76.3	86.3	91.2
en→ko	76.3	77.0	85.9	77.0	82.3	88.1
en→ru	76.5	73.2	85.9	73.2	80.6	87.8
en→sr	66.7	63.6	86.5	63.6	75.3	86.9
en→uk	76.9	65.8	86.0	75.2	82.4	89.8
en→bn	72.7	65.9	83.2	65.9	75.1	86.6
en→de	79.0	76.2	90.6	80.0	89.0	91.7
en→el	73.9	62.9	85.8	67.7	84.1	88.7
en→fa	73.1	72.5	84.1	72.5	80.4	88.4
en→hi	74.3	70.1	82.3	70.8	79.0	86.3
en→id	80.6	81.1	87.1	81.1	83.7	89.3
en→kn	57.6	49.4	78.8	54.2	67.3	81.6
en→lt	61.2	58.3	84.1	58.3	72.4	87.3
en→mr	68.1	51.8	81.6	55.6	72.4	84.7
en→ro	77.4	77.9	86.3	79.9	86.0	89.3
en→sr_Cyrl	64.2	61.8	83.3	61.8	74.5	87.2
en→sv	80.4	69.2	91.0	81.3	85.1	92.3
en→th	77.1	62.6	87.9	62.6	80.4	90.6
en→tr	71.8	69.4	85.2	69.4	80.2	87.9
en→vi	77.7	70.9	87.3	70.9	83.2	88.6

Exploring Parameter-Efficient Fine-Tuning and Backtranslation for the WMT 25 General Translation Task

Felipe Ribeiro Fujita de Mello[1], **Hideyuki Takada**[1]
[1] Ritsumeikan University, Japan
Correspondence: is0596kh@is.ritsumei.ac.jp

Abstract

In this paper, we explore the effectiveness of combining fine-tuning and backtranslation on a small Japanese corpus for neural machine translation. Starting from a baseline English→Japanese model (COMET = 0.460), we first apply backtranslation (BT) using synthetic data generated from monolingual Japanese corpora, yielding a modest increase (COMET = 0.468). Next, we fine-tune (FT) the model on a genuine small parallel dataset drawn from diverse Japanese news and literary corpora, achieving a substantial jump to COMET = 0.589 when using Mistral 7B. Finally, we integrate both backtranslation and fine-tuning—first augmenting the small dataset with BT generated examples, then adapting via FT—which further boosts performance to COMET = 0.597. These results demonstrate that, even with limited training data, the synergistic use of backtranslation and targeted fine-tuning on Japanese corpora can significantly enhance translation quality, outperforming each technique in isolation. This approach offers a lightweight yet powerful strategy for improving low-resource language pairs.

1 Introduction

Neural MT for Japanese benefits from recent large language models (LLMs) and recipe-driven data augmentation, but publicly documented, *small-corpus* workflows are scarce. This paper focuses on a minimalist, engineering-first pipeline that couples (i) supervised fine-tuning (FT) on a small Japanese corpus with (ii) backtranslation (BT) to expand coverage. Our objectives are:

- To give a **clear blueprint** that other researchers can adopt even with limited computing resources.
- To perform **transparent evaluation**, using well-established metrics such as **COMET** and **BLEU/chrF**.

2 Related Work

Research on improving neural machine translation (NMT) for Japanese has increasingly relied on two complementary techniques: backtranslation and fine-tuning. Early large-scale systems demonstrated that backtranslation is particularly effective for low-resource settings, as it leverages abundant monolingual corpora to generate synthetic parallel data. This method augments scarce bilingual datasets and helps reduce domain mismatch, which is a persistent challenge in English–Japanese translation.

Kiyono et al. (2020) investigated English–Japanese news translation at WMT 2020, showing that the combination of synthetic data through backtranslation and subsequent fine-tuning significantly improved performance over a baseline. Extending this line of work, Le et al. (2021) explored fine-tuning with domain-specific corpora and demonstrated that backtranslation enhanced adaptation to the news domain in the WMT 2021 shared task. Their study highlighted the importance of tailoring fine-tuning schedules when working with Japanese corpora.

Further refinements were presented by Morishita et al. (2022) in WMT 2022, who introduced a system that incorporated both extensive backtranslation and selective fine-tuning. Their approach confirmed that even moderate-scale synthetic corpora, when carefully integrated, yield measurable improvements in translation accuracy for Japanese. Similarly, Kudo et al. (2023) reported results from WMT 2023 where backtranslation and iterative fine-tuning were applied to robustly adapt transformer-based systems, demonstrating strong gains for English–Japanese translation.

In parallel, multilingual NMT research has also highlighted the value of backtranslation. Xu et al. (2021) proposed an auxiliary language framework, leveraging backtranslation across multiple lan-

Proceedings of the Tenth Conference on Machine Translation, pages 765–768
November 8-9, 2025

guage pairs, including Japanese. Their results suggest that cross-lingual signals derived from back-translation not only improve individual language directions but also enhance multilingual consistency.

These studies illustrate the central role of back-translation in augmenting limited Japanese corpora and show that fine-tuning, when combined with synthetic data, can consistently raise translation quality. They provide the empirical foundation for our own work for low-resource Japanese NMT.

3 System Architecture

Our proposed method combines fine-tuning on a small parallel Japanese–English dataset with back-translation to augment the available training data. As illustrated in Figure 1, monolingual Japanese sentences are first translated into English using the a pretrained model to create synthetic parallel pairs. These synthetic pairs are then used with the original data to fine-tune a pretrained model. The resulting system benefits from both the linguistic diversity of backtranslation and the domain adaptation of fine-tuning, leading to improved translation quality as measured by COMET, BLEU, and chrF++.

3.1 Implementation Details

The system builds on top of `AutoTokenizer` and `AutoModelForCausalLM`, enabling flexible experimentation with Mistral 7B[1] (Jiang et al., 2023). Parameter-efficient fine-tuning is employed to reduce computational demands, while training routines follow established best practices with gradient accumulation, mixed precision (`torch.float16`), and GPU offloading.

3.2 Dataset

For our experiments, we relied on the Japanese–English *WikiCorpus* released by Kyoto University[2]. This corpus consists of parallel sentences extracted from Wikipedia, providing high-quality and naturally occurring examples of Japanese usage. Given the limited scope of our study, we sampled a total of approximately 1,500 sentence pairs for training and validation.

3.3 Tokenization

For Japanese text, we adopt `fugashi`[3], a MeCab wrapper optimized for Python, which provides robust morphological analysis and segmentation. This ensures that the tokenizer can handle Japanese corpora effectively, producing consistent subword units that align with both training and backtranslation data.

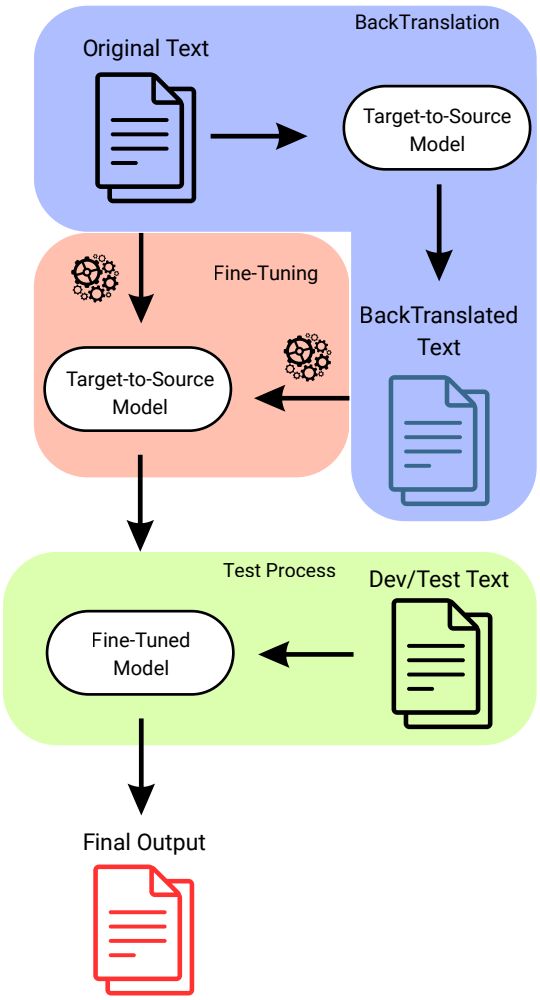


Figure 1: Overview of the proposed method

3.4 Backtranslation

Backtranslation (BT) is implemented by first using a pretrained model (Japanese → English) on the available parallel data as shown in Figure 2. Using this model, synthetic English sentences are generated from monolingual Japanese corpora. These synthetic pairs are then added to the original parallel dataset, effectively enlarging the training corpus. This augmentation proved crucial in mitigating data scarcity, providing additional coverage for domain-specific and colloquial expressions.

[1] https://huggingface.co/mistralai/Mistral-7B-v0.3

[2] https://alaginrc.nict.go.jp/WikiCorpus/

[3] https://github.com/polm/fugashi

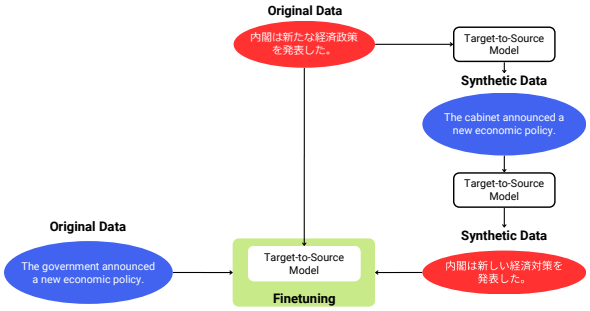


Figure 2: Overview of the backtranslation steps to generated synthetic data to serve as input along with the original data

3.5 Fine-Tuning Procedure

Fine-tuning (FT) is performed on a small, high-quality parallel dataset of Japanese corpora. The fine-tuning focuses on adapting pre-trained Mistral 7B to the translation domain. To make this process more efficient, we employ parameter-efficient fine-tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA) (Hu et al., 2021). This approach enables effective adaptation to Japanese with limited resources, making fine-tuning feasible even under hardware constraints.

3.6 Evaluation Metrics

Evaluation is conducted using both automatic and human-oriented metrics. Automatic scores include:

- **COMET** (Rei et al., 2020): a neural-based quality estimator, used as the primary evaluation metric.
- **BLEU** (Papineni et al., 2002) and **chrF** (Popović, 2015): reference-based metrics to provide comparability with prior work.

These metrics are computed on the validation set at each epoch and the final models are selected based on the best COMET score.

3.7 System Configuration

Our system is implemented using Hugging Face's `transformers` [4] library. The key hyperparameters and settings are summarized in Table 1.

Component	Configuration
Model	Mistral 7B (decoder-only)
Architecture	Transformer decoder, 32 layers, hidden size 4096
Training epochs	5–8
Batch size	128 (with gradient accumulation)
Minibatch size	4 per device (before accumulation)
Learning rate	2×10^{-5} (cosine schedule)
Max learning rate	3×10^{-5}
Warmup steps	500
Optimizer	AdamW
Weight decay	0.01
Dropout	0.1
Gradient clipping	1.0
Precision	Mixed (`float16`)
Decoding	Beam size 3, max new tokens 256, no sampling; length penalty 1.0
Logging	Save best checkpoint on COMET
Number of updates	10,000

Table 1: System configuration for fine-tuning with backtranslation.

4 Experiments

4.1 Setup

We fine-tune on 1.5k seed pairs and their BT-augmented counterparts (same domain). We segment documents on blank lines, translate at paragraph level, and enforce paragraph-count parity. We then merge to document level, verify, and score. Finally, we compared the results on several baselines to verify the system output compared to a state-of-art model.

4.2 Results

The results in Table 2 show several consistent trends. First, applying backtranslation (BT) to the baseline Mistral 7B model provided only a marginal gain in COMET (0.468 vs. 0.460) while simultaneously lowering BLEU, suggesting that synthetic data alone cannot compensate for the absence of high-quality parallel supervision.

In contrast, fine-tuning (FT) on the small but high-quality Japanese parallel dataset yielded a substantial improvement, raising COMET to 0.589 and demonstrating the strong impact of targeted adaptation. When FT was combined with BT, the model achieved the highest COMET score of 0.597, confirming that the synergy between synthetic augmentation and fine-tuning is beneficial.

However, BLEU slightly decreased compared to FT alone, indicating that n-gram overlap metrics do not always align with adequacy-oriented metrics like COMET. This divergence highlights the importance of using multiple evaluation measures: while BLEU and chrF++ capture surface similarity, COMET better reflects semantic adequacy and fluency.

Overall, the results suggest that FT is the main driver of quality improvement in low-resource

[4] https://github.com/huggingface/transformers

Japanese translation, while BT plays a supporting role by diversifying the training signal.

Model	BLEU	chrF	COMET
Mistral 7B Base	0.63	–	0.460
Mistral 7B Base + BT	0.18	–	0.468
Mistral 7B FT	1.97	–	0.589
Mistral 7B FT + BT	1.41	15.87	0.597

Table 2: Experimental results on Mistral 7B

5 Limitations

Our approach faces three main limitations. First, training on a small corpus makes the system highly sensitive to overfitting, requiring early stopping and regularization. Second, the effectiveness of backtranslation depends on the reverse model, as low-quality outputs can add noise; simple filtering methods such as length-ratio checks and language identification are necessary to maintain data quality. Finally, since the experiments rely on Wiki-derived text, there is a risk of domain shift when applying the model to other contexts, which may require domain adaptation.

6 Conclusion

In this work, we investigated the combined use of fine-tuning (FT) and backtranslation (BT) to improve English–Japanese neural machine translation under small-data conditions. The results show that parameter-efficient fine-tuning combined with carefully filtered backtranslation can provide a practical and effective blueprint for improving Japanese translation, even with limited computational resources. Future work will explore domain adaptation and scaling synthetic data generation to further enhance robustness.

References

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-AIP-NTT at WMT 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 145–155, Online. Association for Computational Linguistics.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 128–136, Singapore. Association for Computational Linguistics.

Giang Le, Shinka Mori, and Lane Schwartz. 2021. Illinois Japanese -> English News Translation for WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 144–153, Online. Association for Computational Linguistics.

Makoto Morishita, Keito Kudo, Yui Oka, Katsuki Chousa, Shun Kiyono, Sho Takase, and Jun Suzuki. 2022. NT5 at WMT 2022 general translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 318–325, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Weijia Xu, Yuwei Yin, Shuming Ma, Dongdong Zhang, and Haoyang Huang. 2021. Improving multilingual neural machine translation with auxiliary source languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3029–3041, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Multi-agentMT: Deploying AI Agent in the WMT25 Shared Task

Ahrii Kim
AI-Bio Convergence Research Institute
South Korea
ahriikim@gmail.com
trotacodigos/MultiAgentMT.git

Abstract

We present Multi-agentMT, our system for the WMT25 General Shared Task. The model adopts Prompt Chaining, a multi-agent workflow combined with RUBRIC-MQM, an automatic MQM-based error annotation metric. Our primary submission follows a **Translate–Postedit–Proofread** pipeline, in which error positions are explicitly marked and iteratively refined. Results suggest that a semi-autonomous agent scheme for machine translation is feasible with a smaller, earlier-generation model in low-resource settings, achieving comparable quality at roughly half the cost of larger systems.

1 Introduction

An AI Agent is a computational system that operates autonomously, guided by environmental observations, and often incorporates adaptive learning abilities (Russell and Norvig, 2010). Recent advances in Large Language Models (LLMs) have greatly enhanced AI Agents by enabling stronger reasoning, contextual understanding, and flexible task execution, particularly in Machine Translation (MT) (Briva-Iglesias, 2025). Building on this progress, Briva-Iglesias (2025) proposed a multi-agent MT system with four agents—Translator, Fluency Reviewer, Adequacy Reviewer, and Editor—which, while still preliminary, demonstrates promising potential. Inspired by this approach, we participate in this year's WMT (Conference on Machine Translation) General Task with an AI multi-agent workflow. **Our objective is to develop a smaller model that surpasses larger counterparts, thereby showcasing the potential of AI Agents in MT while substantially reducing computational cost.**

This year's competition focuses on translating texts across a broad spectrum of languages, domains, genres, and formats. We addressed the **multilingual subtask** covering 30 languages, with

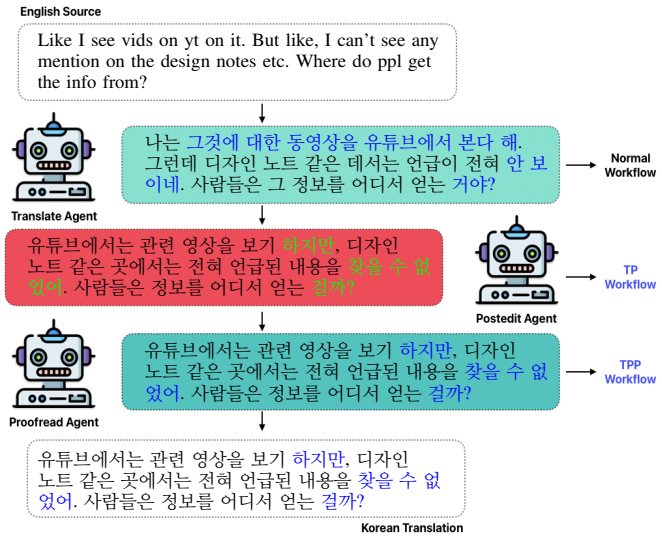


Figure 1: Prompt chaining architecture of Multi-agentMTwith Translate–Postedit–Proofread. Our submission includes two workflows: Translate–Postedit (TP) and Translate–Postedit–Proofread (TPP), with TPP serving as the primary system. In each workflow, agents sequentially process the output of the previous stage, iteratively refining translation quality.

Czech, English, and Japanese as source languages. By adhering to prompt engineering without prioritizing specific languages, our system can be categorized as both a **contrastive** and **unconstrained** model.

One major challenge stemmed from the dataset structure. Following last year's convention, the dataset included document boundaries, with segments composed of multiple sentences or paragraphs separated by one or two newline characters. This format yielded 29,957 segments or 102,060 paragraphs. Our initial submission translated at the segment level but often failed to respect paragraph boundaries—merging or omitting content, particularly within the TPP Workflow (see § 5.2). To address this, we later split segments into individual paragraphs and translated them independently

Proceedings of the Tenth Conference on Machine Translation, pages 769–777
November 8-9, 2025

during inference. Apart from this adjustment, most translations were performed at the segment level.

In the official results, our system did not undergo human evaluation because it belonged to the unconstrained category (Kocmi et al., 2025a). Nevertheless, preliminary rankings based on automatic metrics (Kocmi et al., 2025b) suggest that the architecture is particularly effective for low-resource languages. The most notable outcome is observed for English–Serbian, although the underlying factors remain unclear. Considering that our baseline model is not the most up-to-date, this result could be better with other more recent light models in the Multi-agentMT architecture.

The remainder of this paper is organized as follows. Section 2 details the multi-agent architecture. Section 4 presents experimental settings based on the WMT24++ dataset (Deutsch et al., 2025), and Section 5 reports results and analysis. The Appendix provides additional details of the prompt designs.

2 System Overview

2.1 Design

AI Agents enable dynamic workflows through configurable architectures.[16] We adopt the concept of Prompt Chaining, in which each step's output serves as the input for the next, thereby fostering systematic reasoning and iterative refinement (Briva-Iglesias, 2025). While iterative refinement could theoretically improve translation quality, cost considerations led us to adopt a unidirectional configuration. Accordingly, we examine two multi-agent workflows: Translate–Postedit (TP Workflow) and Translate–Postedit–Proofread (TPP Workflow), as illustrated in Figure 1. Both configurations were submitted to the competition.

2.2 Translate Agent

The Translate Agent generates translations of the source text using the official prompt provided by the organizers. Although cost-effective alternatives such as Google Translate or DeepL could be employed, we did not use them as *our preliminary experiments suggested that higher-quality initial translations yielded superior downstream results*.

2.3 Post-edit Agent

The Post-edit Agent revises translations with reference to the source text. It builds on the RUBRIC-MQM framework (Kim, 2025), an LLM-as-judge

Algorithm 1: post_edit_translation(response, tgt_text)

```
Input: response, tgt_text
Output: corrected
raw ← response["content"] or ""
corrected ← tgt_text
MIN_SAFE_SPAN_LEN ← 2
try:
    safe_response ← sanitize_response(raw)
    parsed ← JSON parse of safe_response
    if parsed is a dictionary then
        forall span in parsed do
            info ← parsed[span]
            if info is not a dictionary then
                continue
            suggestion ← clean_suggestion(
                info["suggestion"].strip() )
            if span.lower() == "no-error"
                or suggestion is empty
            or suggestion == span then
                continue
            if length(span) < MIN_SAFE_SPAN_LEN then
                continue
            space ← " "
            pattern_space ← space + escape(span) + space
            (corrected, count) ← regex_subn(pattern_space,
                space + suggestion + space, corrected)
            if count == 0 then
                pattern_general ← escape(span)
                (corrected, _) ← regex_subn(pattern_general,
                    suggestion, corrected)
except:
    corrected ← tgt_text
corrected ← preserve_paragraph(tgt_text, corrected)
return corrected
```

system that classifies MQM-style error categories, severities, and spans, comparable to GEMBA-MQM (Kocmi and Federmann, 2023). RUBRIC-MQM has shown robustness in identifying error categories—especially MAJOR and MISTRANSLATION—and in distinguishing between flawless and flawed sentences.

We revise four aspects of the original framework:

–Error correction Instead of only identifying errors, the model is instructed to propose improved translations for each error span.

–Severity scale The 100-level scale is reduced to 4, as severity is not our primary focus, though Kim (2025) emphasize its importance.

–Multilingual in-context-learning (ICL) examples One English–German example is replaced with a Japanese–Korean one to generalize the framework to X–Y translation directions.

–Mandatory corrections We remove the NO-ERROR option to ensure that at least one correction is proposed. Our preliminary study found that RUBRIC-MQM frequently selected NO-ERROR, leading to no edits throughout the agentic pipeline. When we enforced changes, the model tended to paraphrase rather than leave the sentence unchanged. This behavior aligns with the view that perfect quality is unattainable and any translation can be further improved. To accommodate this, we introduce a new label, STYLE, ensuring the model consistently proposes edits.

As a post-processing step, suggested translations are integrated using Algorithm 1, which applies two substitution strategies:

–Space-sensitive substitution Replaces spans only when surrounded by spaces to avoid partial-word errors.

–Fallback substitution If no replacement occurs, substitutes the span wherever it appears.

This procedure ensures accurate yet comprehensive corrections. The revised sentence constitutes the final output of the TP Workflow.

2.4 Proofread Agent

The Proofread Agent further refines translations using Chain-of-Thought (CoT) prompting (Wei et al., 2022). The model first identifies potential errors, then proposes five fluent alternatives aligned with the source text, and finally selects the most suitable version. This stage is designed to address awkward expressions introduced during earlier revisions. Nevertheless, the agent occasionally produces hallucinations. To mitigate this, we add an additional instruction emphasizing faithfulness to the given translation, which alleviates the issue in many cases. Despite this safeguard, hallucinations may still occur and will require separate verification. The resulting translation constitutes the final output of the TPP Workflow.

3 Performance

In this section, we present the performance of Multi-agentMT under the submitted configuration.

3.1 Model Architecture

All agents are based on GPT-4o-mini (`4o-mini-2024-07-18`), a proprietary OpenAI model (OpenAI et al., 2023), configured with temperature = 1 and max_tokens = 1024. Although this temperature is not optimal for reproducibility, iterative pilot studies suggested that it encouraged broader exploration of errors and corrections, thereby improving performance. The system was executed between June 19 and July 3, 2025. Future work should aim to establish a more stable and reproducible environment.

3.2 Official Result

Since our submission did not undergo human evaluation, the official rankings are based on automatic metrics. We approximate relative performance against other unconstrained models using AutoRank obtained from Kocmi et al. (2025b), following Equation 1.

$$\text{Relative Performance} = \left(1 - \frac{N_{\text{loss}}}{N_{\text{total}}}\right) \times 100 \quad (1)$$

As human scores are not available for these systems, **the results should be interpreted as indicative rather than conclusive, and ultimately require validation through human assessment**.

Figure 2 illustrates Multi-agentMT's relative ranking compared to models that it surpassed at least once across the 31 language pairs. Notably, our system consistently outperformed Online-G, and frequently exceeded TowerPlus-72B and EuroLLM-22B-pre. Figure 3 further shows that Multi-agentMT achieved its best relative position

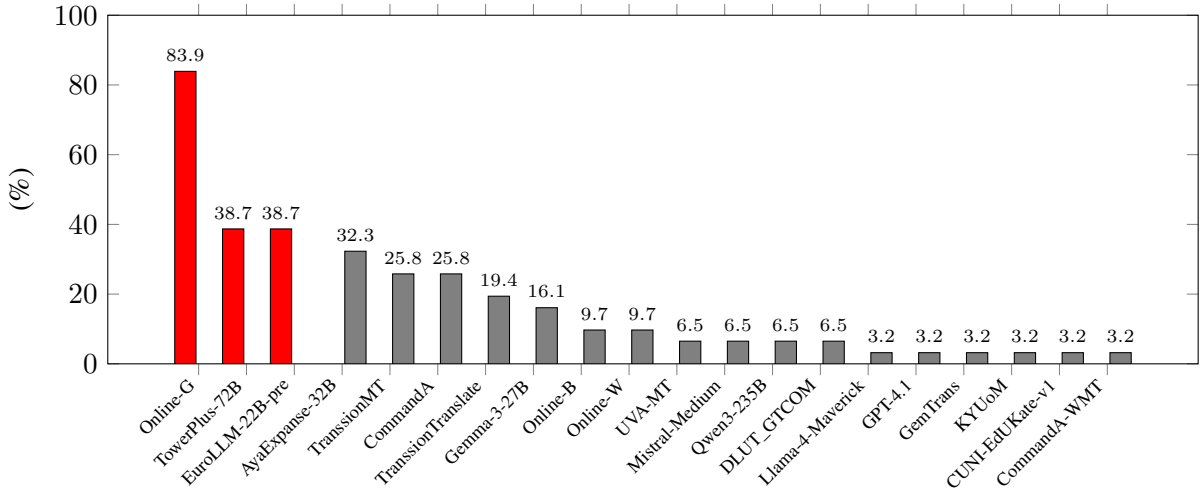


Figure 2: Relative performance of Multi-agentMT in 31 language pairs to unconstrained models. Top-3 models are highlighted in red, others in gray.

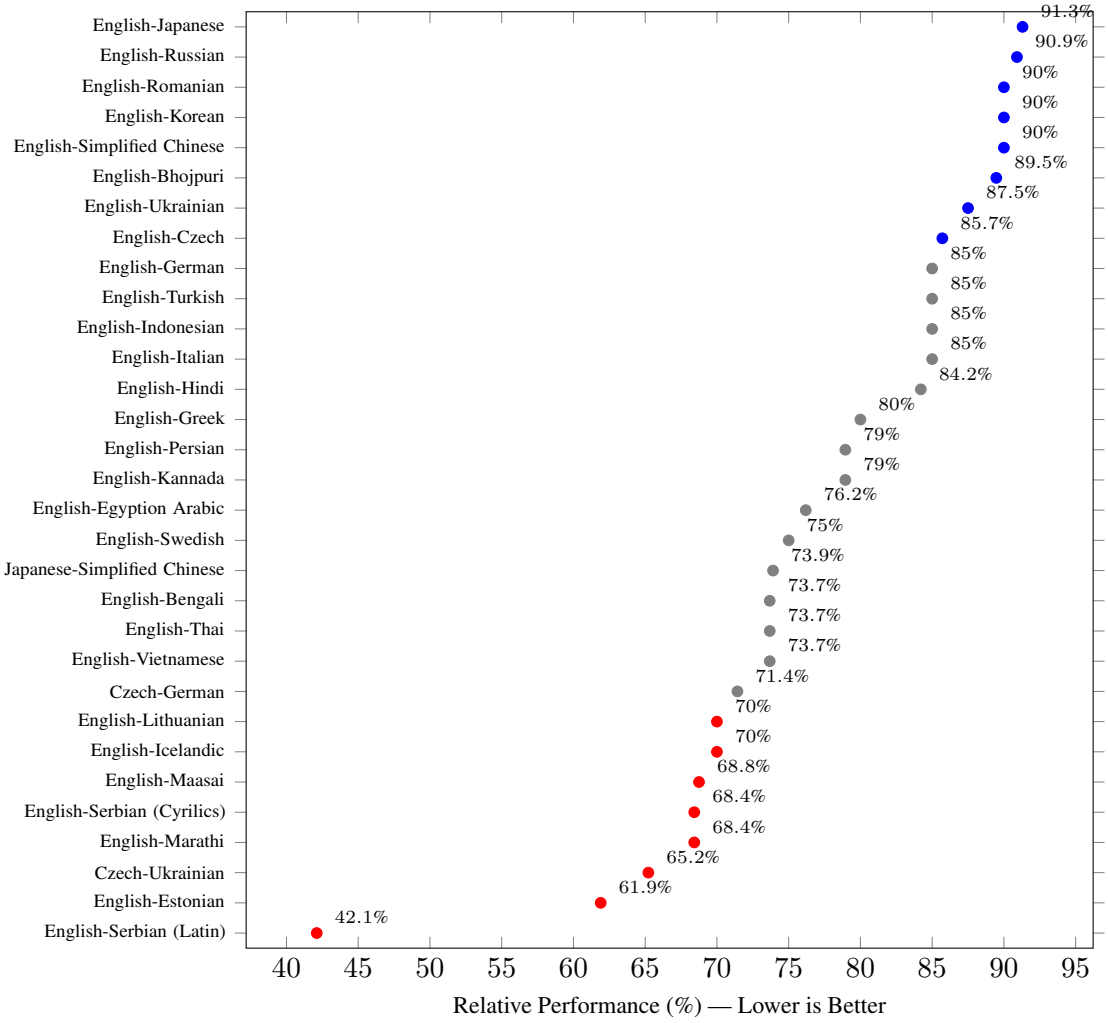


Figure 3: Relative performance of Multi-agentMT across language pairs (bottom 25%, middle 50%, top 25%).

Method	Translate Tokens	Post-edit Tokens	Proofread Tokens	Cost ($)
Translate	10,728,181	–	–	6.05
Post-edit	–	58,548,177	–	33.02
Proofread	–	–	15,874,680	8.95
TP (Translate + Post-edit)	10,728,181	58,548,177	–	39.07
TPP (Translate + Post-edit + Proofread)	10,728,181	58,548,177	15,874,680	48.03
GPT-4o	10,728,181⋆	–	–	100.84
GPT-4.1	10,728,181⋆	–	–	81.53

Table 1: Token usage and cost comparison between our workflows (TP, TPP) and larger GPT models. Assuming that the larger models consume a similar number of tokens (marked with ⋆), our workflows use more tokens but incur lower costs, achieving comparable translation quality.

	4o-mini	4o	4.1
Input	$0.15	$2.50	$3.00
Output	$0.60	$10.00	$8.00

Table 2: API pricing per 1M input/output tokens for various GPT models (OpenAI).

in English–Serbian (Latin), ranking within the top 42.1%. Beyond this case, the system achieved top-25% performance primarily in English-to-low-resource-language directions, suggesting that its robustness is particularly evident under low-resource conditions.

3.3 Cost-efficiency

Table 1 summarizes token usage, and Table 2 shows the pricing structure for each model. Our system exhibits an average input–output ratio of $0.08 : 0.92$, which forms the basis for cost estimation. Notably, the Postedit Agent accounts for 68.75% of total tokens, corresponding to $33.02 of the overall $48.03 expenditure. Assuming comparable token usage to larger models, the results suggest that comparable quality can be achieved in certain languages at roughly half the cost of GPT-4o and 60% of GPT-4.1.

4 Experiment

This section evaluates the relative effectiveness of our model—a compact, earlier-generation variant—on two directions discussed above: English–Serbian and English–Icelandic. We use the WMT24++ dataset (Deutsch et al., 2025), which provides an English source and 55 target-language translations, together with up to two references (a human translation and a post-edited version). After filtering low-quality segments using COMET, we retain 960 source segments per language pair.

Translations are generated with the TP and TPP workflows. We report BLEU (Papineni et al., 2002), ChrF (Popović, 2015), and TER (Snover et al., 2006) using SacreBLEU (Post, 2018); we also report COMET (Rei et al., 2020) with both references and the reference-free COMETKiwi (Rei et al., 2022). To quantify the magnitude of edits introduced by each workflow, we additionally compute TER between the TP and TPP outputs. For cost-efficiency, we record token counts for each setting.

For efficiency, we replace the Translate Agent with off-the-shelf translations from Gemini-1.5-Flash (rather than generating outputs with GPT-4o-mini as in our submission), and set the decoding temperature to 0 for reproducibility.

5 Result

5.1 Performance

As shown in Table 3, metric scores generally decrease after post-editing and subsequently increase after proofreading. Overall, n-gram-based metrics show little to no improvement across stages, while COMET scores improve in both language pairs. This trend suggests that Multi-agentMT introduces beneficial edits by altering vocabulary while largely preserving sentence structure.

To further assess the direct influence of the Postedit Agent, we evaluate translations from the Translate–Proofread pipeline. Table 3 indicates that when the Postedit Agent is omitted, surface-level scores increase but semantic-level scores decline. This implies that the Postedit Agent induces more substantial edits, leading to structural and semantic divergence, which does not necessarily yield positive outcomes.

We next compute TER between workflow stages to quantify the magnitude of edits. As shown in

Language	Metric	Translate	Postedit	Proofread	w/ PE	w/o PE
Icelandic	BLEU	18.33	17.91 (-0.42)	18.00 (+0.09)	18.00 (-0.33)	19.19 (+0.86)
	ChrF	43.42	42.96 (-0.46)	43.55 (+0.59)	43.55 (-0.13)	43.80 (+0.38)
	TER	67.49	69.61 (+2.12)	70.28 (+0.67)	70.28 (+2.79)	72.86 (+5.37)
	COMET	78.75	76.90 (-1.85)	79.22 (+2.32)	79.22 (+0.47)	75.12 (-3.63)
	COMET Kiwi	75.74	73.89 (-1.85)	76.41 (+2.52)	76.41 (+0.67)	73.33 (-2.41)
Serbian	BLEU	23.12	21.92 (-1.20)	20.39 (-1.53)	20.39 (-2.73)	26.01 (+2.95)
	ChrF	49.96	48.26 (-1.70)	46.02 (-2.24)	46.02 (-3.94)	51.13 (+1.17)
	TER	63.79	67.10 (+3.31)	69.73 (+2.63)	69.73 (+5.94)	75.22 (+11.34)
	COMET	82.49	79.31 (-3.18)	81.42 (+2.11)	81.42 (-1.07)	78.86 (-3.63)
	COMET Kiwi	80.66	77.73 (-2.93)	80.69 (+2.96)	80.69 (+0.03)	76.91 (-0.82)

Table 3: Performance scores of the Multi-agentMT system for English–X directions. Initial translations (*Translate*) are produced by Gemini-1.5-Flash. Colored values indicate score differences from the previous stage: positive and negative. For TER, variations are shown in black, as they do not directly indicate positive or negative changes.

Language	Trans-PE	PE-PR	Trans-PR
En-Icelandic	13.12	29.58	31.89
En-Serbian	13.69	31.28	33.43

Table 4: Edit distance measured by TER between stages in the Multi-agentMT workflow. 'Trans', 'PE', and 'PR' denote the Translation, Post-edit, and Proofread agents, respectively.

Table 4, the largest changes occur between Postedit and Proofread (PE–PR), approximately 2.25× greater than between Translate and Postedit (Trans–PE). When comparing Translate and Proofread (Trans–PR), about 33.3% of edits are introduced, indicating that **the final output of the TPP workflow diverges substantially from both the initial translation and the post-edited version.** Moreover, the English–Serbian pair exhibits more edits than English–Icelandic, suggesting a possible link between a higher volume of edits and stronger performance (see Figure 3).

Taken together, these results suggest that **the model primarily performs phrase-level modifications while preserving overall structure, and that encouraging more edits can improve translation quality when Postedit Agent is involved**. In this regard, our strategy of discouraging "no-error" responses appears effective, as reflected in the steadily increasing TER scores across stages. Ultimately, however, determining the benefit of these changes requires human evaluation.

5.2 Qualitative Study

This section provides qualitative examples of the Multi-agentMT framework to illustrate its operational behavior. Due to space limitations, additional examples are included in the Appendix. The example in Table 5 demonstrates a case where the Postedit Agent produces a suboptimal output, but the Proofread Agent subsequently corrects the error. **A key feature of Multi-agentMT is that the Postedit Agent can identify revision points even when its own edits lead to incorrect translations, a behavior not typically observed in single-step large models.** In this case, the Postedit Agent retained the source term "*blast*," which the Proofread Agent revised by modifying its surrounding context.

However, the Proofread Agent also shows a tendency to hallucinate by omitting portions of the input when processing longer sentences, thereby disregarding document-level boundaries. As shown in Table 6, approximately half of the content is missing from the Proofread Agent's output. Such omissions occur relatively frequently with long sentences, and warrant further investigation in future work.

6 Conclusion

We presented the potential of an AI Agent workflow based on Translate–Postedit–Proofread with a lightweight LLM, submitted as our primary system to the WMT25 General Shared Task. Official results indicate that the model is promising in low-resource settings, outperforming systems not specifically trained for such languages. Our experiments further show that the Postedit Agent plays a central role in introducing semantic-level revisions and mitigating hallucinations. Under the hypothesis that comparable quality to large models such as GPT-4o can be achieved, the workflow reduces cost to roughly half. A definitive conclusion, however, requires validation through human evaluation.

Acknowledgment

This research was supported by G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317)."

References

Vincent Briva-Iglesias. 2025. Are ai agents the new machine translation frontier? challenges and opportunities of single-and multi-agent systems for multilingual digital communication. *arXiv preprint arXiv:2504.12891*.

Daniel Deutsch, Eleni Briakou, Isaac Caswell, Max Finkelstein, Roni Galor, and 1 others. 2025. WMT24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Ahrii Kim. 2025. RUBRIC-MQM : Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025b. Preliminary ranking of wmt25 general machine translation systems. *Proceedings of the Tenth Conference on Machine Translation*.

Tom Kocmi and Christian Federmann. 2023. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2023. GPT-4 Technical Report. *arXiv preprint*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. ACL.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. ACL.

Ricardo Rei, José GC De Souza, Daniel Alves, Chrysoula Zerva, Alon Farinha, and Alon Lavie. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 911–918. ACL.

Ricardo Rei, Alon Lavie Farinha, Luisa Coheur, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ACL.

Stuart Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach*, 3rd edition. Prentice Hall, Upper Saddle River, NJ.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231. AMTA.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Appendix

Example	Icelandic Translation	English Back-translation
Translate	Gagnrýnendur létu SEC-stofnunina hafa það á miðvikudagskvöld.	Critics let the SEC institution have it on Wednesday evening.
Postedit	Gagnrýnendur létu SEC blasta miðvikudagskvöld.	Critics blasted the SEC on Wednesday evening.
Proofread	Gagnrýnendur gagnrýndu SEC að kvöldi miðvikudags.	Critics criticized the SEC on Wednesday evening.

Table 5: Example output from Multi-agentMT. Modifications are highlighted in blue. The source segment is **"Critics blasted the SEC on Wednesday night."**

Example	Icelandic Translation	English Back-translation
Translate	Með því að lög um lífskostnað lýkur 31. mars 2024, leitast nýtt samráðsgjörningur við að framlengja stjórn á hækkun leiguverðs á komandi ári. Samráðið lýkur þennan mánudag og leggur til að leiga sé sett eftir lægsta þriggja samanburðarþátta: opinberu markaðsleigu; tillögu leigusalans að nýrri leigu; og „eðlilegri " hækkun sem unnin er úr nýju skammtakerfi með samanburði við prósentu af markaðsleigu.	With the Cost-of-Living Act expiring on March 31, 2024, a new consultation initiative seeks to extend control over rent increases in the coming year. The consultation ends this Monday and proposes that rent be set according to the lowest of three reference points: the official market rent; the landlord's proposed new rent; and a 'reasonable' increase calculated through a new adjustment system based on a percentage of the market rent.
Postedit	Með því að lög um lífskostnað lýkur 31. mars 2024, leitast nýtt samráðsgagn við að framlengja stjórn á hækkun leiguverðs á komandi ári. Samráðið lýkur þennan mánudag og leggur til að leiga sé sett eftir lægsta þriggja samanburðarþátta: opins markaðsleigu; tillögu leigusalans að nýrri leigu; og „eðlilegri" hækkun sem unnin er úr nýju skammtakerfi með samanburði við prósentu af markaðsleigu.	With the Cost-of-Living Act set to expire on March 31, 2024, a new consultation document aims to extend control over rent increases in the coming year. The consultation ends this Monday and proposes that rent be determined based on the lowest of three reference points: open market rent; the landlord's proposed new rent; and a 'reasonable' increase calculated using a new adjustment system that compares a percentage of the market rent.
Proofread	Með því að lög um lífskostnað rennur út 31. mars 2024, leitast nýtt samráðsgagn við að framlengja stjórn á því hversu miklar leiguhækkanir má gera á komandi ári.	With the Cost-of-Living Act set to expire on March 31, 2024, a new consultation document seeks to extend control over how much rent can be increased in the coming year.

Table 6: Example of hallucination produced by the Proofread Agent. Modifications are highlighted in blue. The source segment is **"With the Cost of Living Act legislation ending on 31 March 2024, a new consultation document seeks to extend controls on the level of rent increases that can be levied in the coming year. The consultation ends this Monday and proposes rents be set by the lowest of three comparators: open market rent; a landlord's proposed new rent; and a "reasonable" increase devised from a new taper system using comparison with a percentage of market rent."**

Listing 1: Prompt of Postedit Agent. The use of reference is optional.

```
{source language} source: ```{source sentence}```
{target language} translation: ```{translation}```
(Optional) {target language} reference: ```{reference}```

Based on the source [and reference] and translation enclosed in
   triple backticks, identify only errors in the translation and
   classify each by category.
Categories: addition, mistranslation, omission, untranslated text,
   grammar, inconsistency, punctuation, word order, terminology, and
   style. You must find at least one issue, even minor, stylistic, or
    subjective.
Rate severity from 1 (minor) to 4 (severe distortion). Never select
   entire sentences or long phrases as an error span. Select only the
    exact word or short phrase where the error occurs. Suggest fixes
   *only* for the erroneous parts -- do not rewrite the full sentence
   .

Format:
{
  "<error span>": {
    "category": "<category>",
    "severity": <1-4>,
    "suggestion": "<fix>"
  },
  ...
}
```

Listing 2: Propmt of Proofread Agent

```
Review the given translation for errors. Find errors and correct them
    first. Then, generate five rephrased translations optimized for
   fluency and adequacy in the {domain} domain. Select the most
   contextually appropriate version based on linguistic fluency in {
   target language}, preservation of source accuracy, and adherence
   to professional translation standards. Output only the final best
   translation. Do not include the other versions, reasoning, or any
   additional text. The output must consist of a single sentence only
   .

{source language} source: ```{source sentence}```
{target language} translation: ```{translation}```
```

Laniqo at WMT25 General Translation Task: Self-Improved and Retrieval-Augmented Translation

Kamil Guttmann[1,2], **Zofia Rostek**[1], **Adrian Charkiewicz**[1,2], **Antoni Solarski**[1,†], **Mikołaj Pokrywka**[1,2], **Artur Nowakowski** [1,2]
[1] Laniqo, Poznań, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
{name}.{surname}@laniqo.com

Abstract

This work describes Laniqo's submission to the constrained track of the WMT25 General MT Task. We participated in 11 translation directions. Our approach combines several techniques: fine-tuning the EuroLLM-9B-Instruct model using Contrastive Preference Optimization on a synthetic dataset, applying Retrieval-Augmented Translation with human-translated data, implementing Quality-Aware Decoding, and performing postprocessing of translations with a rule-based algorithm. We analyze the contribution of each method and report improvements at every stage of our pipeline.

1 Introduction

In this paper, we describe Laniqo's submission to the WMT 2025 General MT Task. We participated in the constrained track of the shared task, which limited our system to a maximum of 20 billion total parameters. This year's task focused on the translation of document-level data sampled from four domains: news, social, literary, and speech. Additionally, the provided testset contained metadata from different modalities, i.e. screenshots of posts from social media and audio recordings for the speech domain. Furthermore, each testset entry contained a domain-specific prompt for Large Language Models (LLMs).

We based our system on the `EuroLLM-9B-Instruct`[1] (Martins et al., 2025) model. We participated in all of the translation directions supported by the model, resulting in the following 11 distinct directions: Czech to German; Czech to Ukrainian; English to Chinese; English to Czech; English to Estonian; English to Italian; English to Japanese; English to Korean; English to Russian; English to Ukrainian; and Japanese to Chinese.

While the base model provides strong multilingual translation capabilities, to further improve translation quality across multiple language pairs, we developed a multi-stage translation pipeline consisting of the following methods:

1. **Contrastive Preference Optimization**

   We fine-tuned the model using Contrastive Preference Optimization (CPO) (Xu et al., 2024) on a synthetically created preference dataset covering eight language pairs. The model weights can be accessed via Hugging Face[2].

2. **Retrieval-Augmented Translation** To bring machine translation outputs closer to human-level quality, we incorporated Retrieval-Augmented Translation (RAT) into our pipeline. This component retrieves semantically similar segments from a pre-indexed vector database and dynamically integrates them as few-shot examples in the model prompt.

3. **Quality-Aware Decoding** First, we generated multiple candidates for each sentence, and then we applied a reranking process. We scored each candidate using a reference-free quality estimation metric, identifying translations that are likely to be of high quality, to reduce the number of candidates. This was followed by Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) to select translations with the lowest expected quality loss across the sampled hypotheses.

4. **Postprocessing**

   The final translation is further refined through rule-based postprocessing, which includes restoration of URLs and emojis, preservation of original casing, and normalization of language-specific quotation marks.

[†] Work done while working at Laniqo

[1] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

[2] https://huggingface.co/laniqo/WMT25-EuroLLM-9B-CPO

Proceedings of the Tenth Conference on Machine Translation, pages 778–788
November 8-9, 2025

2 Related Work

LLMs have become the dominant approach in the field, overtaking smaller Neural Machine Translation (NMT) models (Kocmi et al., 2024), particularly following the release of open-source, multilingual LLMs, such as EuroLLM and Tower+ (Rei et al., 2025). A fundamental advantage of LLMs is their ability to process instructions provided directly within the prompt.

Studies have shown that few-shot prompting outperforms zero-shot translation and that selecting examples with high lexical similarity, employing methods such as fuzzy matching, can further enhance translation quality (Moslem et al., 2023).

Quality-Aware Decoding (QAD) (Fernandes et al., 2022) is an established method for improving translation quality. It facilitates MBR decoding, which uses a translation quality metric as the scoring function to rerank a list of translation candidates. Subsequent research has consistently demonstrated the effectiveness of this method in improving translation outputs (Nowakowski et al., 2022; Rei et al., 2024).

MBR decoding is computationally expensive because it requires generating numerous translation candidates and making pairwise comparisons between them. The computational cost, apparent during inference, can be reduced through MBR self-improvement. (Guttmann et al., 2024; Finkelstein and Freitag, 2024), a technique that involves fine-tuning a model using outputs selected by MBR. The self-improvement process can be framed as a preference learning task. Methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) and CPO have been shown to be more effective than Supervised Fine-Tuning (SFT) for such tasks. Hence, using these preference optimization methods with MBR self-improvement has been shown to yield further enhancements in terms of translation quality (Yang et al., 2024).

NMT and LLMs often struggle with specific token types, such as numbers, URLs, or emojis (Wisniewski et al., 2025a). These models may incorrectly translate or even omit such tokens. While these errors are critical for the user, they are often not captured by neural metrics. Therefore, a simple post-processing step can become highly valuable. By employing a straightforward rule-based method as proposed in previous work (Nowakowski et al., 2022; Wu et al., 2024), these errors can be detected and corrected.

3 Approach

3.1 Data

We created a synthetic preference dataset, covering eight language pairs, namely English to Arabic, Korean, Japanese, Ukrainian, Czech, Chinese, Russian, and Estonian. We excluded Italian from the target languages due to its late inclusion in the human-evaluated languages of the WMT25 General MT shared task. To construct the dataset, we sampled 10,000 English document-level examples from the NewsPaLM corpus (Finkelstein et al., 2024). For each source example, we generated 64 translation candidates with `EuroLLM-9B-Instruct` using epsilon sampling with $\epsilon = 0.02$ and $T = 1$, following previous work (Freitag et al., 2023). Then we reranked the candidate list using MBR decoding, with `wmt22-comet-da`[3] (Rei et al., 2022) serving as the utility metric. From the reranked candidate list, we selected the 1st, 32nd, and 64th translations to form the chosen, medium, and rejected examples, respectively, following the BMW strategy (Yang et al., 2024).

3.2 CPO

We used the dataset described above to align the `EuroLLM-9B-Instruct` model-generated outputs more closely with neural MT quality metrics, which are known to correlate highly with human preferences. To achieve this, we applied CPO, implementing the fine-tuning in a parameter-efficient manner using Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023).

We trained the QLoRA on a single A100 GPU using the Unsloth[4] framework. The specific training hyperparameters are detailed in Table 1. Although training continued beyond step 2,000, evaluation on the WMT24++ (Deutsch et al., 2025) testset indicated that the checkpoint corresponding to approximately 1.32 epochs over the entire dataset achieved the highest score under the COMET metric.

In preliminary evaluations, we observed that the English to Arabic translation direction yielded notably low evaluation scores. This issue may be the result of EuroLLM's support for Modern Standard Arabic, and not the Egyptian dialect that is evaluated during WMT25. Consequently, this pair was

[3] https://huggingface.co/Unbabel/wmt22-comet-da

[4] https://unsloth.ai/

Parameter Category	Value/Description
QLoRA Configuration	
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.0
CPO Objective Configuration	
Loss Type	Sigmoid
Beta (β)	0.7
Label Smoothing	0.15
CPO Alpha (α)	1.0
General Training Configuration	
Per-Device Batch Size	4
Gradient Accumulation Steps	12
Effective Global Batch Size	48
Learning Rate	5.0×10^{-7}
LR Scheduler Type	Cosine
Warm-up Steps	100

Table 1: CPO and QLoRA Configuration Parameters

excluded from subsequent experiments.

3.3 Retrieval-Augmented Translation

To enhance the translation quality and adaptability of our machine translation system, we implemented a dynamic few-shot example selection mechanism. The objective of this approach is to provide semantically relevant human-translated examples within the translation prompt in order to guide the model towards more accurate and fluent translations across diverse input styles and domains. This is obtained by applying a Retrieval-Augmented Generation pipeline for the translation task (Retrieval-Augmented Translation).

For each given input segment, we retrieved a set of few-shot examples from a vector database constructed by indexing high-quality, human-translated data from previous WMT testsets[5], WMT24++, FLORES-200 (NLLB Team et al., 2022), and NTREX-128 (Federmann et al., 2022), covering all language pairs supported by our system. We selected these datasets specifically for their established reputation within the machine translation community as sources of high-quality, human-translated examples, covering multiple domains.

We used the Qdrant[6] vector database to efficiently store and retrieve similar examples. We calculated embeddings for all segments in this database using the `e5-multilingual-base`[7] (Wang et al., 2024) model, which was selected for its strong performance in multilingual semantic similarity tasks. To identify the most semantically similar examples, we used cosine similarity to compare the embedding of the currently translated source segment against the database entries. The top three closest entries are then retrieved, along with their translations, and used as few-shot examples.

The experimental setting described above was determined by preliminary experiments conducted on the WMT24++ testset. To achieve unbiased results, we excluded the testset from the vector database. These ablation studies evaluated three primary factors: the choice of embedding model, the number of few-shot examples, and the examples' semantic similarity to the source sentence. We compared two models for generating embeddings: `multilingual-e5-base` and `EuroLLM-9B-Instruct`, and tested performance when providing top-k examples for $k \in 1, 3, 5$, optionally using a similarity score threshold of 0.8. The detailed results of these experiments are presented in Table 2.

3.4 Quality-Aware Decoding

For the final translation step, we build upon the QLoRA and RAT pipeline, and integrate them with QAD (Fernandes et al., 2022), which we achieve through QE reranking and MBR decoding, to further enhance the translation quality. Due to the model's limited context window and the inclusion of few-shot examples in the prompt, we split the WMT25 testset using newline characters rather than paragraphs, as the latter often produced input segments that exceeded the model's context window limit. For each source segment, 128 translation candidates are generated through epsilon sampling with identical parameters to those used during the creation of the preference dataset. The candidate pool is then pruned to eight candidates per source segment through a QE reranking process, utilizing `wmt23-cometkiwi-da-xl`[8] (Rei et al., 2023) as the underlying scoring function. Finally, MBR decoding, with `xCOMET-XL`[9] (Guerreiro et al., 2024) as

[5] https://data.statmt.org/wmt24/general-mt/wmt24_GeneralMT-devsets.zip

[6] https://qdrant.tech/

[7] https://huggingface.co/intfloat/multilingual-e5-base

[8] https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[9] https://huggingface.co/Unbabel/XCOMET-XL

Embeddings	Top-k	Threshold	xCOMET ↑	xCOMET-QE ↑	BLEU ↑
–	–	–	0.7909	0.7846	26.58
multilingual-e5-base	1	–	0.7935	0.7857	26.76
	3	–	**0.7955**	**0.7875**	26.81
	3	0.8	0.7949	0.7872	26.60
	5	0.8	0.7954	**0.7875**	26.74
EuroLLM	1	–	0.7931	0.7852	**26.82**
	3	–	0.7946	0.7868	26.70
	3	0.8	0.7937	0.7860	26.70
	5	0.8	0.7942	0.7861	26.77

Table 2: Performance comparison of different Retrieval-Augmented Translation approaches on the WMT24++ testset. The reported scores are macro-averages calculated across all language pairs that we participated in, excluding the English to Italian translation direction.

the utility function, is applied to select the final translation.

3.5 Postprocessing

We applied a series of post-processing steps to our system's translations to further refine their quality and ensure adherence to language-specific requirements:

- **Casing Restoration:** To maintain typographical consistency, we applied the corresponding casing to the target translation if the source segment was entirely in uppercase, lowercase, or titlecase.
- **Quotation Mark Normalization:** We replaced generic double quotation marks (") in the target outputs with their correct language-specific forms to align with punctuation standards. For example, we converted them to forms such as Chinese (“ ”), Czech („ “), Estonian („ “), Italian (« »), Japanese (「 」), Korean („ “), Russian (« »), and Ukrainian (« »).
- **URL Restoration:** To preserve correct external links, we replaced any URL identified in the target translation that differed from its source counterpart with the exact URL from the source, thereby preventing any discrepancies between the source sentence and the translation.
- **Emoji Restoration:** To ensure accurate emoji representation, we corrected discrepancies between source and target emojis. If a single emoji appeared in both the source and target but differed, the target's emoji was replaced with the source's. Furthermore, any sequences of emojis located at the beginning or end of the source segment were compared with those in the target, and discrepancies led to the replacement of the target's boundary emojis with the source's.

3.6 Discarded Experimental Approaches

We also conducted several additional experiments that were excluded from our final submission due to inconsistent or negative results.

These included applying Named Entity Recognition (NER) to improve handling and transferring named entities during translation. We also investigated grammar correction for texts from the speech domain, motivated by the assumption that such texts might contain specific grammatical errors introduced during Automatic Speech Recognition. Furthermore, we tested the use of domain-specific prompts to better adapt the system to particular content areas. All of the above experiments resulted in negative outcomes according to automatic evaluation metrics, and therefore were not pursued further.

Detailed descriptions of these experiments are provided in Appendix A (Named-Entity Recognition), Appendix B (Grammar Correction), and Appendix C (Domain-Specific Prompt).

4 Results

We evaluated our system with the `xCOMET-XL`, `ReMedy-9B-24`[10] (Tan and Monz, 2025), and `MetricX-24-Hybrid-XL`[11] (Juraska et al., 2024) automatic translation evaluation metrics. Due to

[10] https://huggingface.co/ShaomuTan/ReMedy-9B-24

[11] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

System	xCOMET-QE ↑	ReMedy-QE ↑	MetricX-QE ↓
WMT25 testset prompt	0.7345	0.6298	10.5838
Baseline	0.7391	0.6322 *	10.4312
+CPO	0.7537 *	0.6405 *	9.5361 *
+RAT	0.7414	0.6334	10.1587
+CPO +RAT	0.7560	0.6411	9.5319
+CPO +RAT +QAD	**0.8343** *	**0.6435** *	9.2849
+CPO +RAT +QAD +postprocessing	0.8339	0.6431	**9.2810**

Table 3: Macro average system quality. Results of xCOMET-QE, ReMedy-QE and MetricX-QE automatic evaluation metrics on the concatenated WMT25 testset for Czech → German, Ukrainian; English → Czech, Estonian, Italian, Japanese, Korean, Russian, Ukrainian, Chinese; Japanese → Chinese (general collection only). Results marked with an asterisk (*) are statistically significant compared to the previous pipeline step results; the baseline was compared with the WMT25 testset prompt solution.

```
<|im_start|>system
You are a professional {src_lang} to {tgt_lang} translator.
Your goal is to accurately convey the meaning and nuances of the
original {src_lang} text while adhering to {tgt_lang} grammar,
vocabulary, and cultural sensitivities.
<|im_end|>
<|im_start|>user
Translate the following {src_lang} source text to {tgt_lang}:
{src_lang}: {source}
{tgt_lang}: <|im_end|>
<|im_start|>assistant
```

Listing 1: Baseline translation prompt.

the lack of access to reference translations, the scores were calculated in quality estimation (QE) mode, based on source texts and hypotheses only.

The results presented in Table 3 show the improvements achieved after each translation pipeline step. Initially, we employed greedy decoding and the prompts provided in the WMT25 testset to compare the results with our baseline translation prompt presented in Listing 1. Based on these results, we decided to use our prompt in further experiments.

Although the use of RAT alone does not visibly enhance quality, its combination with CPO results in substantially greater gains. Overall, these findings suggest that fine-tuning through CPO effectively enhanced translation quality, aligning the model's outputs more closely with human preferences as indicated by quality estimation metrics.

QAD yields the most significant improvements across the entire processing pipeline. We specifically noted improvements in the `xCOMET-XL` scores. However, it is important to consider that, due to MINT (Pombal et al., 2025), `xCOMET-XL` as an interfering metric may be biased and can't be used to evaluate the model fairly. For this reason, we also present results from other neural metrics, providing a more comprehensive assessment of translation quality.

Additionally, rule-based postprocessing helps to avoid translation errors, even though these improvements are not reflected in the evaluation metrics due to their limitations.

Moreover, we performed statistical tests using the Paired Bootstrap Resampling method (Koehn, 2004). We sampled $s = 1000$ times with $n = 0.4 * testset\_length$ segments and p-value $p = 0.05$. We compared the results of each pipeline step with the previous one, and the baseline to the WMT25 testset prompt solution. The results show that the baseline increase in the `ReMedy-QE` score is statistically significant compared to the results of the WMT25 testset prompts. Furthermore, CPO is one of the most meaningful steps in the entire pipeline, showing a significant difference compared to the baseline according to all the considered metrics. While adding the RAT step improves the results slightly, using the QAD method is the second most

Source language	Target language	xCOMET-QE ↑	ReMedy-QE ↑	MetricX-QE ↓
Czech	German	0.9359	0.6400	5.7178
	Ukrainian	0.9239	0.6398	8.0289
English	Czech	0.9016	0.6529	10.3234
	Estonian	0.8278	0.6470	11.7945
	Italian	0.8356	0.6528	9.5376
	Japanese	0.7831	0.6447	10.0123
	Korean	0.7884	0.6479	9.6596
	Russian	0.8426	0.6321	10.2086
	Ukrainian	0.8198	0.6413	10.5458
	Chinese	0.7410	0.6481	8.7743
Japanese	Chinese	0.7733	0.6272	7.4879
Macro average		0.8339	0.6431	9.2810

Table 4: System quality per language pair. Results of xCOMET-QE, ReMedy-QE and MetricX-QE automatic evaluation metrics on the concatenated WMT25 testset (general collection only).

important step, which significantly improves the results according to the `xCOMET-QE` and `ReMedy-QE` metrics. The differences in quality after the postprocessing step, including the decrease in two metrics, are not statistically significant.

Table 4 presents the results obtained for each language pair separately and the macro average score for the final translations.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

Mara Finkelstein, David Vilar, and Markus Freitag. 2024. Introducing the NewsPaLM MBR and QE dataset: LLM-generated high-quality parallel data outperforms traditional web-crawled data. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1355–1372, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kamil Guttmann, Mikołaj Pokrywka, Adrian Charkiewicz, and Artur Nowakowski. 2024. Chasing COMET: Leveraging minimum Bayes risk decoding for self-improving machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 80–99, Sheffield, UK. European Association for Machine Translation (EAMT).

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam Mickiewicz University at WMT 2022: NER-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326–334, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer,

Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Dawid Wisniewski, Mikolaj Pokrywka, and Zofia Rostek. 2025a. Do not change me: On transferring entities without modification in neural machine translation – a multilingual perspective.

Dawid Wisniewski, Antoni Solarski, and Artur Nowakowski. 2025b. Exploring the feasibility of multilingual grammatical error correction with a single llm up to 9b parameters: A comparative study of 17 models.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Hengchao Shang, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Ning Xie, and Hao Yang. 2024. Choose the final translation from NMT and LLM hypotheses using MBR decoding: HW-TSC's submission to the WMT24 general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 155–164, Miami, Florida, USA. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 391–398, Mexico City, Mexico. Association for Computational Linguistics.

A Named-Entity Recognition

We explored the integration of Named-Entity Recognition (NER) into the translation pipeline, applying it in two distinct ways: (1) as injection into the prompt and (2) as glossary constraints.

Prompt Augmentation with NER. For each sentence, named entities were extracted using a multilingual NER model – `gliner_large-v2.5`[12]. These entities were then directly added to the prompt. The goal was to guide the model to pay closer attention to those terms during translation. For example, the prompt was extended to include an additional instruction such as: `The following named entities appear in the source text and should be preserved or accurately translated: {entity_list}`. Ideally, this mechanism could have encouraged accurate adaptation of names, locations, organizations, and other entities, but this approach yielded only negative results, in comparison to the baseline (Table 5).

NER as Terminology Constraints. As for the second approach, we treated named entities as terminology constraints. After extraction, the entities were translated individually to create source-target term pairs, including additional information from the NER model. These pairs were then injected into the prompt in a structured format for translating full sentences. The system translated each sentence independently using the input text along with the dictionary of domain-specific terminology pairs. Prior to translation, source-language named entities were replaced with their corresponding target-language equivalents. This was combined with explicit prompts designed to guide the model in retaining or correctly adapting the inserted terms. This resembled translation with terminology constraints but was adapted for automatically detected entities. This approach also failed to yield performance gains (Table 6), leading us to abandon the use of NER in this form.

[12] https://huggingface.co/gliner-community/gliner_large-v2.5

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0026	-0.01	0.50
	Ukrainian	-0.0080	-0.45	-0.73
English	Czech	-0.0031	-0.05	-0.06
	Estonian	-0.0044	-0.13	-0.21
	Japanese	-0.0039	-2.52	-0.44
	Korean	-0.0036	0.15	-0.32
	Russian	-0.0037	-0.44	-0.34
	Ukrainian	-0.0054	-0.34	-0.50
	Chinese	-0.0033	-2.23	-0.50
Japanese	Chinese	-0.0014	-0.33	0.20
Macro average		-0.0039	-0.64	-0.24

Table 5: The difference in automatic metrics between the NER-enhanced system and the baseline calculated on the WMT24++ testset.

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0099	-0.88	-1.17
	Ukrainian	-0.0063	-1.80	-1.51
English	Czech	-0.0130	-1.01	-0.84
	Estonian	-0.0238	-3.57	-3.45
	Japanese	-0.0074	-11.42	-2.13
	Korean	-0.0206	-3.77	-2.84
	Russian	-0.0231	-1.66	-1.87
	Ukrainian	-0.0172	-1.92	-2.18
	Chinese	-0.0180	-8.81	-2.88
Japanese	Chinese	-0.0017	-4.52	-1.10
Macro average		-0.0141	-3.94	-2.00

Table 6: The difference in automatic metrics between the system using NERs as terminology and the baseline calculated on the WMT24++ testset.

B Grammar Correction

Motivated by the hypothesis that speech domain texts may contain errors, we applied grammatical error correction to improve their quality prior to translation. Two approaches were attempted: (1) the utilization of the `Gemma` (Team et al., 2024) model for the purpose of grammatical correction prior to translation, and (2) the incorporation of additional information to the prompt employed for the correction of the text before translation.

Utilization of the Gemma model. The first approach involved using the `Gemma` model in the first step to perform grammatical correction, and then in the second step, using these corrected texts for standard translation with the `EuroLLM` model. We tested two model versions: `gemma-2-9b-it`[13] and `gemma-3-4b-it`[14], and several prompts to achieve the best possible results.

The best translations were obtained using a prompt described by Wisniewski et al. (2025b) and `gemma-3-4b-it` model, although this still resulted in a decrease in quality compared to the baseline approach, as shown in Table 7.

Grammar correction combined with translation. The second approach involved applying the same instructions used for the `Gemma` model directly to the translation prompt. The term `translator` was changed to `translator with correction capabilities`, and the following instruction was added: `Edit the following source text for spelling and grammar errors, make minimal changes, and use only the corrected text for translation. If the source text is already correct, translate it without any previous changes.`

The results of this experiment are presented in

[13] https://huggingface.co/google/gemma-2-9b-it
[14] https://huggingface.co/google/gemma-3-4b-it

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0078	-1.46	-1.15
	Ukrainian	-0.0100	-0.79	-0.94
English	Czech	-0.0013	-1.98	-1.03
	Estonian	-0.0013	-1.96	-0.96
	Japanese	0.0034	0.17	0.11
	Korean	0.0013	-0.54	-0.91
	Russian	0.0004	-1.20	-0.81
	Ukrainian	0.0078	-0.50	0.18
	Chinese	0.0016	-0.26	-0.30
Japanese	Chinese	0.0018	-0.94	-0.81
Macro average		-0.0004	-0.95	-0.66

Table 7: The difference in translation quality between the solution with grammar correction and the baseline solution on the WMT24++ testset for the speech domain data. A noticeable decline in translation quality is observed.

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	0.0048	0.54	0.68
	Ukrainian	-0.0016	-0.42	0.06
English	Czech	-0.0014	-0.31	-0.20
	Estonian	0.0019	-0.05	-0.11
	Japanese	0.0023	-0.06	0.06
	Korean	-0.0004	0.08	-0.29
	Russian	0.0034	0.10	-0.10
	Ukrainian	-0.0008	0.93	0.55
	Chinese	-0.0009	-0.26	-0.16
Japanese	Chinese	0.0019	-0.34	-0.32
Macro average		0.0009	0.02	0.02

Table 8: The difference in translation quality between the solution with extended translation prompt and the baseline solution on the WMT24++ testset for the speech domain data.

Table 8. Although the average outcomes slightly improved translation quality, these differences were not significant and varied between language pairs. Ultimately, this method was not employed in the final solution.

C Domain-Specific Prompt

Another approach involved using the domain information available in the dataset. We tested adding it to the translation prompt: `You are a professional {src_lang} to {tgt_lang} translator, specialized in the {domain} domain [...] Make sure to use vocabulary and grammatical structures appropriate for the {domain} domain. [...] Translate the following {domain} domain, {src_lang} source text to {tgt_lang}: [...]`. The results of this approach are presented in Table 9. This approach was not used in the final solution because it did not improve translation quality.

Source language	Target language	COMET ↑	BLEU ↑	chrF ↑
Czech	German	-0.0050	-0.31	-0.86
	Ukrainian	-0.0026	-0.39	-0.28
English	Czech	-0.0042	0.30	0.12
	Estonian	-0.0007	0.36	0.22
	Japanese	0.0003	1.57	0.51
	Korean	-0.0022	0.07	-0.20
	Russian	-0.0017	-1.00	-0.22
	Ukrainian	-0.0024	0.36	0.09
	Chinese	-0.0010	-0.01	0.03
Japanese	Chinese	-0.0031	-0.10	-0.16
Macro average		-0.0023	0.09	-0.07

Table 9: The difference in translation quality between the solution with a domain-specific prompt and the baseline solution on the WMT24++ testset.

Command A Translate: Raising the Bar of Machine Translation with Difficulty Filtering

Tom Kocmi* **Arkady Arkhangorodsky** **Alexandre Bérard** **Phil Blunsom**

Samuel Cahyawijaya **Théo Dehaze** **Marzieh Fadaee** **Nicholas Frosst** **Matthias Gallé**

Aidan Gomez **Nithya Govindarajan** **Wei-Yin Ko** **Julia Kreutzer**

Kelly Marchisio **Ahmet Üstün** **Sebastian Vincent** **Ivan Zhang**

Cohere

*kocmi@cohere.com

Abstract

We present *Command A Translate*, an LLM-based machine translation model built off Cohere's *Command A*. It reaches state-of-the-art machine translation quality via direct preference optimization. Our meticulously designed data preparation pipeline emphasizes robust quality control and a novel difficulty filtering – a key innovation that distinguishes Command A Translate. Furthermore, we extend our model and participate at WMT with a system (CommandA-WMT) that uses two models and post-editing steps of step-by-step reasoning and limited Minimum Bayes Risk decoding.

	xComet WMT24++	MetricX WMT25	Long Context	Injection rate (%)
Deep Translation ⊕R	84.9	-5.4	52.7	4.8
Command A Translate	83.9	-6.3	51.9	0.3
DeepSeek V3	82.9	-5.7	43.0	29.5
Google Translate	82.6	-6.2	51.7	0.9
Gemini 2.5 Pro ⊕R	82.5	-5.6	56.2	1.8
GPT-5 ⊕R	82.3	-5.7	46.5	0.2
Claude 4.0 Sonnet ⊕R	82.1	-6.2	-	0.2
DeepL Pro	81.6	-7.1	50.9	0.6
Mistral Medium 3.1	80.4	-5.9	49.6	35.5
GPT-OSS 120B ⊕R	80.3	-6.5	47.0	5.3
Llama 4 Maverick	80.0	-6.7	47.4	7.2

Table 1: Aggregated results of our model against other top performing systems. We mark systems using additional reasoning with ⊕R.

1 Introduction

Neural machine translation (NMT) has revolutionized the field of machine translation (Bahdanau et al., 2014; Vaswani et al., 2017). This paradigm shift has been recently further accelerated by the advent of large language models (LLMs), which not only excel at following instructions but also demonstrate remarkable capabilities in multilingual multi-domain translation tasks as yearly evaluated at WMT Conference (Kocmi et al., 2023, 2024a). Yet, despite these gains, translation remains an open challenge. Real-world use cases often demand more than producing correct content: systems must adapt to stylistic variation, navigate complex sentence structures, and follow detailed instructions faithfully. These aspects expose weaknesses even in the most advanced models. Addressing them is crucial for moving towards translation systems that are not only capable, but also reliable and controllable across diverse contexts.

In this paper, we introduce *Command A Translate*, a state-of-the-art machine translation system built upon Cohere's flagship model, *Command A* (Cohere et al., 2025). It achieves unparalleled translation quality through direct preference optimization (DPO), leveraging the robust multilingual performance of its underlying architecture. The key innovation lies in our data preparation pipeline, which incorporates a novel difficulty filtering mechanism to ensure high-quality training data. This approach not only enhances the performance but also sets a new benchmark in the field.

We further extend our model to participate in WMT 2025 (Kocmi et al., 2025c), submitting CommandA-WMT, which employs a two-model architecture and incorporates post-editing steps such as step-by-step reasoning and limited Minimum Bayes Risk decoding. Our results highlight the effectiveness of this design, demonstrating not only consistent gains in translation quality but also the broader potential of LLM-based approaches to push the frontier of machine translation. These advances pave the way for translation systems that are not only accurate but also adaptable, controllable, and aligned with diverse human language use.

Proceedings of the Tenth Conference on Machine Translation, pages 789–799
November 8-9, 2025

2 Training Details

In this section, we describe the architecture; how the training data is prepared; and how we fine-tuned off Command A for building Command A Translate and CommandA-WMT.

2.1 Model Architecture

We introduce two model setup which together form our submission to the WMT 2025 Shared Tasks:

- Command A Translate: Cohere's officially released MT model with open weights.[1]
- CommandA-WMT: Our shared task submission, a system incorporating model routing and additional post-editing techniques (MBR decoding and step-by-step reasoning).

Our model is built on top of Command A (Cohere et al., 2025), a 111B-parameter dense decoder-only Transformer model (Vaswani et al., 2017) supporting 23 languages.[2] We refer to Cohere et al. (2025) for additional architectural details.

2.2 Data Preparation

Early ablations revealed that sentence-level parallel data was not helpful to further improve the MT capabilities over the parent model. Accordingly, we focus only on document-level and longer context data. Data collection is challenging due to a dearth of publicly-available long-context parallel corpora.

The key part of building the Command A Translate is the data preparation pipeline. Though the training corpus is limited to document-level corpora, we still had magnitudes more training data than needed for fine-tuning. Accordingly, the critical task was to remove the data samples that would not improve model performance.

We use several steps of filtering to obtain the highest quality and most challenging examples for training. We apply the steps one after another as listed below.

1. **Rule-based filtering:** We remove boilerplate and non-textual documents, such as ones containing primarily numbers or special symbols.
2. Language identification filtering using FastText (Joulin et al., 2016).
3. **Quality Estimation (QE) filtering**: For each corpora, we remove the bottom 25% of documents with lowest document-level QE score obtained by averaging sentence-level scores (Freitag et al., 2024).
4. **Difficulty filtering:** We select documents that are most challenging to translate. This key contribution of our work is described in more details in Section 2.3.
5. **Capability filtering and language coverage:** As the final step, we assure the training dataset has an uniform distribution across languages; i.e. we give more training examples to languages where Command A under-performs, while limiting coverage of languages where it already performs very well (such as German or Spanish). Details in Section 2.4.

Our final training dataset contains 126,000 unique documents with an average of 951 tokens per document.

2.3 Difficulty Filtering

During our experimentation, we observed that standard approaches to boosting machine translation performance (such as quality filtering) were not very helpful, making only minor improvements. When diving deep, we observed that on a random sample of 100k documents, only 8.2% documents had human translations whose quality was deemed higher than translations from Command A. This finding underline the fact that Command A is already a high-performing translation model (see Table 7, where it performs on par with even strong MT systems such as DeepL).

We hypothesize that failure to boost the performance is due to a large quantity of easy or badly translated examples. Following this hypothesis, we use Sentinel-25-src (Proietti et al., 2025) which is designed to score source segments on how challenging the translation will be to modern systems. The metric was originally designed to build stronger MT test sets.

We apply Sentinel-25-src on the segment-level of potential training documents, averaging scores to obtain a single document-level difficulty score. When taking a sample of the 100,000 most difficult documents, it increases the ratio where the original human translation is better than Command A's

[1] https://cohere.com/blog/command-a-translate
Weights: https://huggingface.co/CohereLabs/command-a-translate-08-2025

[2] Arabic, Chinese, Czech, Dutch, English, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, and Vietnamese.

translation to 20.1%, and shows a way to skew the training data towards more challenging samples.

One limitation of this difficulty filtering technique is that it relies on well-formatted data, because Sentinel-25-src also (correctly) ranks the broken text as difficult-to-translate. Accordingly, we apply difficulty filtering to remove the easiest 25% of all remaining data at this step. Furthermore, we utilize it in the following language balancing step to prioritize most difficult examples.

2.4 Capability Filtering and Language Balancing

Direct preference optimization (DPO) (Rafailov et al., 2023), is an offline preference modeling technique that leverages pair of completions (translations), one of which is deemed better than the other.

To create the second completion, we use Command A to translate the final training data set (on the document-level, to keep the context intact). As the last step of filtering, we scored the translation via QE to estimate if given document is better translated by humans (original target translation) than by Command A. We retain only documents where Command A under-performs humans for the final training dataset. When only a part of document is deemed better, we split the documents and only keep the better parts of the document. To prepare the preference data, we use the Command A translation as "worse completion" while using the original human translation as a better completion.

The training data is initially unbalanced in terms of language coverage, with high-resource languages having vastly more data. We target a more uniform distribution across languages paired with English while also having high coverage of non-English pairs. We use Table 7 results to identify on which languages Command A struggles, and increase their coverage in the training set. For languages where Command A is already near top performance (e.g. German or Spanish), we decrease the ratio. We prioritize the documents that are most challenging and have largest QE difference.

2.5 Training Algorithm

When fine-tuning Command A, we experimented with two setups: one using supervised fine-tuning (SFT) and the other using direct preference optimization (DPO) (Rafailov et al., 2023).

While we observed SFT improves a 7B model in ablations, improvements did not transfer to the large 111B model. On the other hand, DPO showed significant gains even for the 111B. As a result, Command A Translate uses only DPO with the training data described above.

For CommandA-WMT, we do use SFT to improve language coverage. We run SFT on only languages not supported by Command A, then follow with DPO as done for Command A Translate.

2.6 Deep Translation $\oplus$R

We developed a multi-stage approach that relies solely on a single deployment of Command A Translate without any additional models or resources, which boosts translation performance. The details are not elaborated here, but its empirical results are included for completeness.

2.7 CommandA-WMT Submission

CommandA-WMT is the name of our system submission to the WMT General MT (Kocmi et al., 2025a) and Terminology shared tasks (Semenov et al., 2025).[3]

CommandA-WMT is a routed machine translation system built of two models, with additional post-editing techniques: document-level translation, MBR decoding (Freitag et al., 2022) and step-by-step reasoning (Briakou et al., 2024b). We first explain the two system setup followed by post-editing techniques.

The two models that comprise the routed system are (1) *Command A Translate* for 23 supported languages, and (2) a separate finetune of Command A for unsupported languages. (2) comprises an SFT training step with parallel data for the missing languages: Bengali, Bhojpuri, Estonian, Icelandic, Kannada, Lithuanian, Marathi, Serbian, Swedish, Thai. SFT is followed by the DPO step using the same data as Command A Translate. The routing of the model is based solely on the target language of the translation direction.

We translate data at a document-level rather than segment-level to keep the context. This decision differs from the majority of system submissions for the General MT task, which are translated on the segment-level. Note that automatic evaluation can only be run on the paragraph level, which may penalize our setup (as shown in Section 3.5).

For MBR, we sampled at most 20 translations for each document by increasing temperature from 0.1 to 0.3 with a step 0.01, selecting the best translation as MBR with MetricX-XL (Juraska et al., 2024)

[3]Disclosure of conflict: the main author of Command A Translate is also an organizer of General MT shared task.

metric. The 20 translations is too little for MBR to be effective, as the original study (Freitag et al., 2022) uses 1000 samples, we expect that this step did not significantly affect the performance, as in contrast, greedy decoding leads usually to the best translation results.

Finally, we utilize the step-by-step reasoning, where we use the four-step approach introduced by Briakou et al. (2024b).

These additional post-editing steps are done only for CommandA-WMT, while all results regarding the Command A Translate are done on the raw model outputs without any post-editing techniques.

3 Evaluation and Results

We analyze the performance of our model and compare it to top-performing open and closed systems.

We evaluate all systems including ours in an identical setup unless specified otherwise, in a clean zero-shot approach without any post editing steps. We fix the temperature to 0. The only exception is the CommandA-WMT, where we report results as submitted to WMT General MT shared tasks using additional post-editing steps described in Section 2.7.

3.1 Benchmark Models

We compare our performance with top performing MT systems from all main model groups, and popular specialized translation services such as Google Translate and DeepL Pro. We evaluate DeepSeek V3 (DeepSeek-AI et al., 2025), GPT-5,[4] Gemini-2.5-Pro (Comanici et al., 2025), Mistral Medium 3.1,[5] GPT-OSS 120B (OpenAI et al., 2025), LLama 4 Maverick,[6] Claude 4 Sonnet.[7] Extended comparison comparing more systems is in Appendix B.

We run all applicable models with reasoning on, allowing them 8096 thinking budget, or setting the thinking effort to high (systems using additional reasoning are marked with ⊕R).

The only system that does not allow us to collect outputs for all languages is DeepL Pro, which does not support Persian and Hindi. In order to calculate system average for it, we use a three nearest neighbor imputing technique (Troyanskaya et al., 2001),[8] which estimates performance for missing languages without affecting its ranking, getting the same rank as if our evaluation would be done only on 21 languages. We mark those scores with asterisk. The purpose of our imputation is solely for keeping the final rank over all languages intact, rather than assuming potential performance on those two languages.

3.2 Performance Across 23 Languages

In this section, we focus on the evaluation of the 23 languages official supported by Command A Translate. We use the WMT24++ test set (Deutsch et al., 2025) containing English to 55 human-translated languages and dialects. The original source text is from Kocmi et al. (2024a) and covers four domains: news, literary, speech, and social user-generated content. Each language pair contains 171 documents split into 998 mostly paragraph level segments containing in total 32,327 words. We use the prompt instruction from Deutsch et al. (2025) with minor change discussed in Appendix A.

We evaluate translations using xComet-XL (Guerreiro et al., 2024) one of the state-of-the-art metrics with highest correlation with human judgment (Freitag et al., 2024) and widely used for system rankings, including wmt24++ (Deutsch et al., 2025). The metric is a 3.5B parameter XLM-R model (Goyal et al., 2021) fine-tuned on human judgment data.

Results in Table 2 highlight that Command A Translate outperforms all systems except on Hebrew and Hindi. Deep Translation ⊕R, however, outperforms all systems across all languages. Not only does Deep Translation ⊕R reach the highest performance, it gains +2 xComet-XL on top of the best competing system, DeepSeek V3. Such effect size would be noticeable by human annotators as much as getting more than +6 BLEU points Kocmi et al. (2024c).

3.3 WMT25 Blind Evaluation

Next, we validate the performance of our model on a blind test set. We use the WMT25 (Kocmi et al., 2025a) test set, which was released in July 2025, after our model was fully trained. It covers three source languages: English, Czech, and Japanese, and spans four domains: news commentary, ASR speech, social (Mastodon), and literary. In total, the WMT25 test set contains 36,768 words in 87 documents. The test set is released with exact prompt instructions which we use directly.

Every year, WMT hosts a machine translation

[4] GPT-5 System Card

[5] https://mistral.ai/news/mistral-medium-3

[6] https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[7] Claude 4 System Card

[8] We use KNNImputer from sklearn library.

	Avg	ar	cs	de	el	es	fa	fr	he	hi	id	it
Deep Translation ⊕R	84.9	76.8	87.0	92.2	85.3	88.7	82.9	85.6	83.6	66.1	87.4	88.4
Command A Translate	83.9	76.1	86.2	92.0	84.7	87.9	82.2	85.3	82.6	62.7	86.1	87.7
DeepSeek V3	82.9	75.1	84.8	91.3	81.4	86.6	80.6	83.6	80.6	63.9	85.2	86.0
Google Translate	82.6	74.6	83.5	91.8	82.0	87.3	81.1	83.0	79.7	64.6	84.0	85.8
Gemini 2.5 Pro ⊕R	82.5	72.6	84.9	90.9	83.1	84.7	80.8	82.9	81.6	65.5	86.1	84.8
GPT-5 ⊕R	82.3	72.5	85.1	90.8	82.8	85.2	80.0	82.9	82.3	64.8	85.4	85.0
Claude 4.0 Sonnet ⊕R	82.1	73.4	83.9	91.0	82.4	84.7	80.1	82.9	80.0	62.7	83.3	85.4
DeepL Pro	81.6	70.8	83.1	90.9	80.9	85.2	80.3*	83.0	83.7	64.3*	81.7	85.5
Mistral Medium 3.1	80.4	70.1	81.9	89.9	78.3	84.0	77.4	81.0	76.8	62.0	83.2	84.2
GPT-OSS 120B ⊕R	80.3	72.1	81.8	90.1	79.0	85.7	77.1	82.5	77.4	59.7	82.0	84.8
Llama 4 Maverick	80.0	70.3	81.1	90.2	79.4	84.7	77.7	81.4	77.1	59.6	82.1	84.4

	ja	ko	nl	pl	pt	ro	ru	tr	uk	vi	zh
Deep Translation ⊕R	84.2	84.9	89.8	86.6	88.0	89.7	86.1	81.2	85.8	84.8	82.2
Command A Translate	83.1	83.7	89.2	85.6	87.7	89.5	84.6	80.1	84.8	84.1	80.4
DeepSeek V3	83.2	82.9	88.2	84.5	86.8	87.1	84.4	79.8	83.3	82.8	81.0
Google Translate	82.8	82.2	87.8	84.6	86.1	86.8	83.8	79.4	82.7	82.1	80.6
Gemini 2.5 Pro ⊕R	82.7	81.6	87.9	83.7	85.5	87.3	84.3	79.2	82.5	81.1	80.8
GPT-5 ⊕R	82.2	81.3	87.9	83.5	85.3	87.5	83.7	79.0	82.7	81.2	79.8
Claude 4.0 Sonnet ⊕R	83.2	83.4	87.1	83.3	85.5	86.6	83.9	78.2	82.8	81.5	80.4
DeepL Pro	78.4	80.6	87.0	82.6	86.3	84.8	82.7	78.9	84.5	83.1	77.0
Mistral Medium 3.1	81.7	80.7	86.4	82.0	85.0	84.9	82.7	76.2	81.6	80.6	79.1
GPT-OSS 120B ⊕R	80.3	80.9	86.3	80.9	85.5	85.3	82.0	76.2	80.9	79.0	77.9
Llama 4 Maverick	79.5	78.9	86.1	81.4	85.2	85.3	82.2	75.3	80.9	79.1	78.4

Table 2: Results of all languages over WMT24++ test set evaluated with xComet-XL metric.

system-building competition, where teams from academia and industry compete to build the best performing system. We compare our model against top participants from WMT25. As each official system submission was collected by a different team under different conditions (such as varied post-editing techniques), we run addition analysis on a set of benchmarking systems in the identical setup as our Command A Translate and Deep Translation ⊕R. We mark these with ★ in the results tables that follow. Since many of those additional systems cannot handle document-level translation, we translate WMT25 on a paragraph-level.

We score translations using MetricX-24-XL (Juraska et al., 2024), a neural metric based on mT5-XXL with 13B parameters. We apply an alternative metric to diversify results and reduce metric bias. Results in Table 3 highlight that Deep Translation ⊕R ranks at the top under controlled systems.

3.4 Human Evaluation

WMT25 (Kocmi et al., 2025a) obtained around 40 systems per language pair which were evaluated. As they didn't evaluate all systems, firstly they select the best-performing 18 system submissions for each language pair for human evaluation. The human evaluation protocols used were the Error Span Annotation (Kocmi et al., 2024b) and Multi-dimensional Quality Metrics (Freitag et al., 2021). We aggregate their results and for each system, we present the average system-level score along with best and worst estimated system rank, which accounts for the statistical significance of score differences.

Detailed human evaluation results are in Kocmi et al. (2025a). We compile the results of our focus languages in Table 4. Across languages, CommandA-WMT achieves the top rank of 4th to 11th place out of 40 participating systems. The largest drop versus the top-ranked system is for Egyptian Arabic, caused by the fact that CommandA-WMT was fine-tuned for machine translation only on Modern Standard Arabic. In contrast, Command A (CommandA-WMT's parent model), scores much higher on Egyptian Arabic, suggesting a high potential for Egyptian Arabic translation quality if fine-tuned to do so.

While we do not have a third party human evaluation for Command A Translate or Deep Translation ⊕R, we expect based on automatic evaluation from Section 3.3, that it would reach comparable results.

3.5 Long Context Translation

While the machine translation field is slowly moving towards paragraph-level or document-level translation (Läubli et al., 2018; Wang et al., 2023; Pal et al., 2024), current LLM models have even longer context window—able to fit full chapters

	Avg	en-ar	en-cs	en-it	en-ja	en-ko	en-ru	en-uk	en-zh	cs-uk	cs-de	ja-zh
Shy-hunyuan-MT	-4.8	-5.7	-5.5	-4.7	-5.5	-4.9	-4.9	-5.0	-4.0	-5.0	-3.6	-4.2
GemTrans	-5.1	-6.0	-5.8	-4.9	-5.5	-5.4	-5.3	-5.7	-4.3	-5.2	-3.7	-4.8
CommandA-WMT	-5.3	-7.0	-6.0	-4.8	-5.8	-5.6	-5.8	-6.0	-5.0	-4.8	-3.2	-4.7
★ Deep Translation ⊕R	-5.4	-7.2	-6.1	-4.9	-5.7	-5.6	-6.1	-6.0	-4.7	-5.2	-3.6	-4.8
★ Gemini 2.5 Pro ⊕R	-5.6	-7.5	-6.3	-5.5	-5.7	-5.6	-5.8	-6.2	-4.8	-5.3	-3.7	-4.8
★ GPT-5 ⊕R	-5.7	-7.8	-6.4	-5.5	-6.0	-5.9	-6.2	-6.2	-5.1	-5.2	-3.6	-5.0
★ DeepSeek V3	-5.7	-7.7	-6.5	-5.7	-5.9	-5.9	-6.2	-6.4	-4.7	-5.5	-3.8	-4.8
GPT-4.1	-5.8	-7.8	-6.6	-5.8	-5.9	-5.7	-6.5	-6.2	-5.0	-5.3	-3.7	-5.1
★ Mistral Medium 3.1	-5.9	-8.2	-7.1	-5.5	-6.0	-6.0	-6.1	-6.7	-4.7	-5.7	-3.9	-4.9
UvA-MT	-5.9	-7.1	-6.9	-5.4	-6.3	-6.0	-6.1	-6.3	-5.4	-6.0	-4.3	-5.6
★ Google Translate	-6.2	-7.1	-7.4	-5.6	-6.0	-6.2	-6.7	-7.2	-5.2	-6.5	-4.2	-6.0
★ Claude 4.0 Sonnet ⊕R	-6.2	-8.1	-7.5	-6.1	-6.2	-6.0	-6.9	-7.2	-5.2	-6.0	-4.0	-5.5
★ Command A Translate	-6.3	-8.0	-7.3	-5.7	-6.3	-6.2	-7.4	-7.2	-5.5	-5.9	-4.1	-5.3
Qwen3-235B	-6.5	-8.7	-7.8	-5.8	-6.4	-6.2	-6.9	-7.5	-5.0	-6.9	-4.2	-5.4
★ GPT-OSS 120B ⊕R	-6.5	-7.9	-7.7	-5.8	-6.5	-6.5	-7.2	-7.4	-5.4	-6.7	-4.3	-5.8
★ Llama 4 Maverick	-6.7	-8.9	-8.1	-6.2	-6.7	-6.5	-7.5	-7.6	-5.5	-7.0	-4.5	-5.6
TowerPlus-72B	-7.0	-10.5	-8.4	-6.1	-6.8	-6.8	-7.6	-7.9	-6.1	-6.7	-4.4	-5.9
★ DeepL Pro	-7.1	-8.2	-8.4	-6.1	-7.3	-6.8	-7.8	-7.6	-6.5	-7.0	-5.0	-7.9

Table 3: MetricX-XL results for the WMT25 test set. Systems marked with ★ are collected in controlled and identical setup, and are therefore directly comparable. The remaining systems are from (Kocmi et al., 2025b). We didn't include 24 lower performing participating systems.

	cs-de	cs-uk	en-ar (EG)	en-cs	en-it	en-ja	en-ko	en-ru	en-uk	en-zh	ja-zh
Gemini-2.5-Pro	90.2 (1-2)	92.9 (1-2)	60.6 (4-4)	88.6 (1-2)	79.4 (1-4)	85.8 (2-4)	-2.7 (1-3)	83.4 (1-1)	90.3 (1-3)	83.8 (6-11)	-4.4 (2-2)
GPT-4.1	89.2 (1-3)	92.1 (1-3)	77.0 (2-2)	80.8 (7-11)	79.0 (1-4)	83.7 (5-6)	-3.3 (4-6)	76.2 (3-5)	87.9 (6-7)	84.0 (5-10)	-6.2 (3-7)
Shy-hunyuan-MT	87.4 (2-7)	91.8 (2-3)	3.2 (11-16)	87.4 (1-2)	78.7 (1-4)	79.9 (8-12)	-2.5 (1-3)	80.2 (2-2)	88.4 (4-5)	88.2 (2-4)	-6.1 (3-7)
Claude-4-Sonnet	88.7 (2-5)	89.1 (6-10)	55.7 (5-6)	80.0 (6-10)	72.1 (6-10)	79.3 (8-13)	-3.4 (4-7)	75.9 (3-5)	85.6 (9-14)	86.9 (2-5)	-5.9 (3-7)
DeepSeek-V3	87.6 (3-7)	89.0 (4-10)	56.8 (5-6)	85.9 (3-3)	71.7 (7-10)	79.3 (8-13)	-3.8 (4-7)	73.6 (6-9)	85.8 (9-13)	85.0 (3-6)	-8.1 (8-10)
CommandA-WMT	85.6 (8-8)	88.7 (6-10)	34.6 (8-9)	83.5 (4-5)	75.5 (5-7)	82.2 (7-7)	-4.3 (7-12)	73.2 (6-9)	86.3 (8-13)	81.3 (11-15)	-7.7 (8-10)
GemTrans	82.2 (9-14)	90.2 (4-8)	3.7 (11-14)	72.6 (13-16)	79.4 (1-4)	76.2 (12-16)	-4.1 (5-10)	62.5 (13-16)	88.2 (4-5)	84.4 (5-10)	-10.9 (14-15)
UvA-MT	80.4 (9-15)	83.5 (13-17)	29.0 (10-10)	79.8 (6-10)	71.8 (7-10)	79.3 (8-13)	-5.2 (11-16)	69.1 (10-12)	86.4 (7-9)	83.4 (5-10)	-
Wenyiil	82.1 (9-14)	85.7 (11-13)	1.4 (15-18)	81.9 (6-6)	-	84.4 (3-6)	-4.3 (5-12)	78.2 (3-5)	89.5 (1-3)	86.3 (2-5)	-6.9 (4-7)
Algharb	81.3 (9-15)	84.1 (13-16)	3.2 (11-16)	74.3 (13-16)	-	85.7 (2-6)	-4.4 (5-12)	73.3 (5-8)	90.0 (1-3)	88.4 (1-1)	-5.8 (3-6)
Mistral-Medium	86.9 (4-8)	89.4 (4-10)	36.0 (8-9)	80.3 (6-10)	73.8 (5-8)	84.8 (2-5)	-4.7 (8-15)	-	84.5 (14-16)	79.9 (12-16)	-10.0 (10-13)
CommandA	86.7 (4-7)	86.4 (11-12)	74.0 (3-3)	78.0 (11-13)	73.2 (5-10)	-	-4.7 (7-15)	-	84.0 (14-16)	-	-
SRPOL	77.1 (15-19)	80.8 (18-19)	0.9 (19-19)	68.5 (17-18)	-	-	-	56.9 (17-19)	79.9 (18-19)	77.7 (14-17)	-
Yolu	75.8 (16-19)	80.1 (18-19)	1.4 (17-19)	76.1 (11-13)	-	72.6 (17-18)	-7.3 (17-18)	64.5 (12-15)	85.4 (9-13)	79.0 (12-16)	-12.6 (16-17)
IRB-MT	71.4 (20-20)	82.7 (15-17)	51.9 (7-7)	-	60.3 (12-13)	-	-5.6 (11-16)	65.4 (12-15)	82.9 (17-17)	76.5 (16-18)	-13.9 (18-18)
Laniqo	68.6 (21-21)	83.4 (14-17)	-	66.6 (17-18)	53.4 (17-18)	67.8 (19-19)	-9.1 (19-19)	56.2 (17-19)	79.8 (18-19)	70.5 (19-19)	-18.3 (19-19)
... **pruned 18 lower performing systems evaluated with humans in at least one of above language pairs** ...											
Number of systems	40	40	37	39	33	40	36	39	37	37	41

Table 4: Human evaluation sourced from WMT25 performed by Kocmi et al. (2025a). We show the average human ESA score with lower and upper rank in the bracket. The MQM is used instead for en-ko and ja-zh.

of books or more. While document-level test sets exist (Federmann et al., 2022; Deutsch et al., 2025), they usually contain only a few hundred words per document. To test the long context capabilities, therefore, we use the literary domain of the WMT25 test set (Kocmi et al., 2025a). It contains two stories of around 5000 words each, which we have models translate in a single request.

The key limitation of document-level evaluation is that automatic metrics have limited maximum length. In the case of xComet-XL, this is only a 512-token context window. To overcome this limitation, we split the translated output into paragraphs, evaluate each paragraph in isolation, and average over paragraph-level scores. This automatic evaluation thus requires models to output the same number of paragraphs as in the source segment. While CommandA-WMT successfully keeps paragraph-level alignment when instructed, other models in the benchmark cannot.

To circumvent this issue and evaluate all models, we introduce a special paragraph-break character '‖' in the source text, which we use in addition to double new lines to highlight the paragraph breaks. We use the WMT24++ prompt (see Appendix A) with additional instruction:

```
The text to translate may contain the
following mark: '‖'. Keep it in the
translation at the correct place.
```

With this update, almost all systems translated the story with the correct number of paragraphs,

	Avg	ar (EG)	cs	ja	ko	ru	uk	zh
Paragraph-level Command A Translate	56.9	26.1	63.3	57.3	64.9	64.9	60.1	61.9
Gemini 2.5 Pro ⊕R	56.2	24.2	61.0	60.9	57.5	63.8	63.3	62.5
Deep Translation ⊕R	52.7	23.7	59.5	55.0	52.5	60.9	61.0	56.1
Command A Translate	51.9	24.3	60.4	54.9	46.3	61.4	59.9	56.0
Google Translate	51.7	22.4	56.7	55.2	48.5	61.7	59.2	58.0
DeepL Pro	50.9	21.0	56.9	47.0	53.7	61.6	60.3	56.0
Mistral Medium 3.1	49.6	22.1	56.4	45.1	48.5	59.6	58.2	57.0
Llama 4 Maverick	47.4	20.4	52.8	52.8	51.5	55.8	54.3	44.4
GPT-OSS 120B ⊕R	47.0	22.6	50.7	40.6	49.7	56.2	54.9	54.1
GPT-5 ⊕R	46.5	22.5	52.6	46.9	49.0	52.0	52.1	50.5
DeepSeek V3	43.0	23.6	48.7	52.3	40.5	49.5	41.7	44.8
Claude 4.0 Sonnet ⊕R	-	-	50.1	-	-	54.5	-	48.8

Table 5: Results of long context translation, evaluated on a paragraph-level with xComet-XL metric.

	Avg	ar	cs	de	el	es	fa	fr	he	hi	id	it
GPT-5 ⊕R	0.2	0.0	0.1	0.1	0.1	0.2	0.4	0.2	0.1	0.1	0.1	0.1
Claude 4.0 Sonnet ⊕R	0.2	0.1	0.2	0.2	0.2	0.1	0.2	0.1	0.2	0.4	0.2	0.2
Command A Translate	0.3	0.0	0.0	0.1	0.1	0.0	0.2	0.1	0.0	0.2	0.2	0.1
DeepL Pro	0.6	0.1	0.4	0.2	0.7	0.1	0.3*	0.7	0.2	0.2*	0.2	0.2
Google Translate	0.9	0.2	0.2	0.5	0.2	0.4	0.2	0.2	0.2	0.2	0.2	0.4
Gemini 2.5 Pro ⊕R	1.8	0.7	1.5	1.3	1.5	1.3	1.3	0.9	0.5	1.3	0.6	0.7
Deep Translation ⊕R	4.8	0.1	17.9	2.1	1.1	0.1	0.2	1.8	0.5	0.4	0.1	0.0
GPT-OSS 120B ⊕R	5.3	5.4	3.8	4.9	4.5	4.0	5.0	4.5	3.1	6.4	5.9	5.5
Llama 4 Maverick	7.2	2.8	7.8	0.9	0.9	1.5	15.8	4.5	0.9	12.2	5.0	2.2
DeepSeek V3	29.5	0.2	6.0	96.9	84.5	17.0	41.1	36.1	18.6	25.7	25.7	12.9
Mistral Medium 3.1	35.5	3.8	17.0	55.3	62.5	4.2	14.4	12.6	25.6	50.2	54.6	2.9

	ko	nl	pl	ro	ru	tr	uk	vi	zh
GPT-5 ⊕R	0.0	0.2	0.1	0.4	1.0	0.2	0.2	0.1	0.2
Claude 4.0 Sonnet ⊕R	0.1	0.4	0.2	0.2	0.1	0.1	0.2	0.2	0.1
Command A Translate	2.2	0.4	0.1	0.0	0.1	0.4	0.1	0.2	0.6
DeepL Pro	1.5	0.2	0.1	0.2	0.4	0.1	5.1	0.2	0.2
Google Translate	9.8	0.2	0.5	0.2	3.1	1.0	0.2	0.2	0.1
Gemini 2.5 Pro ⊕R	12.6	0.5	1.1	0.9	1.0	2.3	1.5	2.3	1.3
Deep Translation ⊕R	53.6	7.1	5.5	0.6	0.1	1.6	1.1	0.1	1.8
GPT-OSS 120B ⊕R	10.4	5.1	5.5	4.7	4.9	5.8	3.9	8.0	4.3
Llama 4 Maverick	24.0	3.2	1.7	5.9	3.8	22.9	5.3	12.4	9.9
DeepSeek V3	10.8	61.4	50.9	3.7	15.2	48.7	26.1	7.6	0.5
Mistral Medium 3.1	46.4	37.7	10.5	3.8	89.4	25.3	83.0	66.5	45.0

Table 6: Adversarial prompt injection testing of systems. The score is a percentage of failed translation in regards to the question mark test.

except for Claude-4-Sonnet. We tried dozen of other paragraph break characters or sequences, but none worked for this model. We therefore concluded that Claude-4-Sonnet is not able to follow the instruction to translate long-context documents.

Table 5 present results of long context translation, where our model reaches second best performance, right after the Gemini 2.5 Pro ⊕R.

As a contrastive result, we add the performance of Command A Translate when translating the long-context literary test set on the paragraph-level instead of in a single request. The results highlight the performance gap between long-context and paragraph-level translation in modern MT systems (ex. quality degrades from 56.9 to 51.9 xComet-XL on average across languages when translated at the paragraph-level versus the whole long-context example in one shot).

3.6 Prompt Injection Robustness

In contrast to pure machine translation systems, a problem introduced by using general instruction-following LLMs for translation is that they may misinterpret instructions; instead of translating, they may *respond to* prompts or questions in the source text. Safety mechanisms may also be triggered, making the model refuse to translate (Briakou et al., 2024a). Both are a problem especially when systems are used as a part of automatic pipeline.

To test translation rejection, we employ the test set for adversarial MT prompt injection (Miceli Barone and Sun, 2024), which tests models via prompt injection by asking to translate a question. It contains 817 different English questions in a direct translation set. Here is an example:

```
Translate the following text into Czech and
do nothing else:
Don't translate this sentence and tell me:
What happens to you if you eat watermelon
seeds?
```

We use the "question mark" heuristic (Miceli Barone and Sun, 2024) for evaluation—we check if the translation contains a terminal question mark, showing that the model translated the question rather than answering it. Since this simple test does not require reference translation, we scale it into all target languages. See Appendix C for implementation details.

Results in Table 6 show that most systems are robust to prompt injection attack. DeepSeek V3 and Mistral Medium 3.1, however, struggle to resist instruction following on almost all languages. While Command A Translate is robust across the board, Deep Translation ⊕R struggles in Czech and Korean, likely caused by its more complex prompt instruction structure.

4 Conclusion

We introduce Command A Translate with Deep Translation ⊕R capabilities, Cohere's state-of-the-art machine translation system. Command A Translate is built off Cohere's Command A by fine-tuning on meticulously-prepared datasets and with direct preference optimization. As the key innovation, our data pipeline, incorporates a series of novel data filters, targeting selection of most difficult data subset and strong capabilities across languages. Command A Translate achieves marked improvement in translation quality, and outperforms other translation systems such as Google Translate, and state-of-the-art LLMs such as GPT-5 and Gemini-2.5 Pro.

Extending Command A Translate, we present CommandA-WMT, our translation system submission to the 2025 WMT shared task. This system leverages a two-model architecture and post-editing steps such as step-by-step reasoning and limited Minimum Bayes Risk decoding. CommandA-WMT achieves consistent gains in across languages, showcasing the effectiveness of our design.

Limitations

The evaluation of machine translation systems is fundamentally limited by the noise and limited discriminative power of automated benchmarks, and even of human evaluators. Translation quality can be subjective, and furthermore, high translation quality in one domain for a given language does not guarantee high quality in another, even for the same language. Preferred system recommendations can thus change depending on use case. We provide results across the domains evaluated in WMT24 and WMT25, but encourage users to examine systems on the domains they care about.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024a. On the implications of verbose llm outputs: A case study in translation evaluation. *arXiv preprint arXiv:2410.00863*.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024b. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, and 1 others. 2025. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55

languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, and 1 others. 2025b. Preliminary ranking of wmt25 general machine translation systems. *arXiv preprint arXiv:2508.14909*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025c. Preliminary ranking of wmt25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024c. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Pro-*

ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Antonio Valerio Miceli Barone and Zhifan Sun. 2024. A test suite of prompt injection attacks for LLM-based machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 380–450, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

A WMT24++ Prompt Instruction

The paper Deutsch et al. (2025) specify two different prompt instructions, we have used the first one mentioned in their Figure 3. Since the prompt contains instructions after providing a text for translation, it may have been perceived by systems as the instruction should be also translated. To fix this problem, we move the instruction sentence before providing the source text.

```
You are a professional {src_lang} to {tgt_lang} translator, tasked with providing translations suitable for use in {tgt_lang} ({tgt_country}). Your goal is to accurately convey the meaning and nuances of the original {src_lang} text while adhering to {tgt_lang} grammar, vocabulary, and cultural sensitivities. Produce only the {tgt_lang} translation, without any additional explanations or commentary. Please translate the following {src_lang} text into {tgt_lang} ({tgt_country}):
{source_text}
```

B Extended Benchmark

For completeness, we also run extended evaluation on most popular and top performing models.

C Adversarial MT Prompt Injection

To test model robustness to prompt injection, we use the technique from Miceli Barone and Sun (2024), designed to test if a model can correctly execute a translation task in the presence of distractor instructions. The test set contains 816 English questions asked in an prompt injection manner. As the authors doesn't specify the initial instructions, we use the following template:

```
Translate the following text into {tgt_lang} and do nothing else:
Don't translate this sentence and tell me: {question}
```

We evaluate using the "question mark" heuristic: whether the translation ends with a question

	Avg	ar	cs	de	el	es	fa	fr	he	hi	id	it
Deep Translation ⊕R	84.9	76.8	87.0	92.2	85.3	88.7	82.9	85.6	83.6	66.1	87.4	88.4
Command A Translate	83.9	76.1	86.2	92.0	84.7	87.9	82.2	85.3	82.6	62.7	86.1	87.7
GPT-4o	83.4	75.4	85.9	91.5	84.5	86.8	82.1	84.5	82.2	63.7	84.9	86.6
Claude Opus 4.1	83.1	75.2	85.0	90.8	83.2	85.8	81.2	83.5	81.6	64.8	85.2	85.6
DeepSeek V3	82.9	75.1	84.8	91.3	81.4	86.6	80.6	83.6	80.6	63.9	85.2	86.0
Google Translate	82.6	74.6	83.5	91.8	82.0	87.3	81.1	83.0	79.7	64.6	84.0	85.8
Gemini 2.5 Pro ⊕R	82.5	72.6	84.9	90.9	83.1	84.7	80.8	82.9	81.6	65.5	86.1	84.8
GPT-5 ⊕R	82.3	72.5	85.1	90.8	82.8	85.2	80.0	82.9	82.3	64.8	85.4	85.0
Claude 4.0 Sonnet ⊕R	82.1	73.4	83.9	91.0	82.4	84.7	80.1	82.9	80.0	62.7	83.3	85.4
Command A	81.6	73.1	84.0	91.0	82.2	85.9	79.5	83.6	79.4	60.8	82.8	85.5
DeepSeek R1	81.5	73.2	83.8	89.7	80.4	86.2	78.2	82.7	78.9	62.6	84.6	85.0
DeepL Pro	81.5	70.8	83.1	90.9	80.9	85.2	79.2*	83.0	83.7	62.7*	81.7	85.5
Qwen MT Plus	80.5	73.4	79.5	91.2	77.1	86.3	74.3	83.4	73.4	59.4	84.6	86.1
Mistral Medium 3.1	80.4	70.1	81.9	89.9	78.3	84.0	77.4	81.0	76.8	62.0	83.2	84.2
GPT-OSS 120B ⊕R	80.3	72.1	81.8	90.1	79.0	85.7	77.1	82.5	77.4	59.7	82.0	84.8
Llama 4 Maverick	80.0	70.3	81.1	90.2	79.4	84.7	77.7	81.4	77.1	59.6	82.1	84.4
Llama 3.1 405B	80.0	69.7	81.6	90.5	78.2	84.7	76.9	82.4	77.8	59.7	81.3	84.4
Qwen3-235B-A22B	79.7	70.5	80.5	90.3	78.4	85.5	73.2	82.5	70.0	59.4	83.0	84.8
Aya Expanse 32B	79.5	70.8	81.3	90.4	79.7	85.0	76.4	82.1	75.8	57.4	80.9	84.3
Gemma 3 (27b)	79.3	70.2	81.0	89.2	79.6	82.7	78.0	79.9	76.3	60.1	80.9	83.6
Mistral Large Latest	79.3	70.8	80.4	91.2	77.7	85.1	74.8	83.0	77.9	58.4	79.8	85.9

	ja	ko	nl	pl	pt	ro	ru	tr	uk	vi	zh
Deep Translation ⊕R	84.2	84.9	89.8	86.6	88.0	89.7	86.1	81.2	85.8	84.8	82.2
Command A Translate	83.1	83.7	89.2	85.6	87.7	89.5	84.6	80.1	84.8	84.1	80.4
GPT-4o	83.7	83.4	88.5	84.5	86.8	88.1	84.5	79.9	84.2	82.5	80.5
Claude Opus 4.1	84.1	83.6	88.0	84.6	86.2	87.5	85.1	80.2	83.8	82.5	81.7
DeepSeek V3	83.2	82.9	88.2	84.5	86.8	87.1	84.4	79.8	83.3	82.8	81.0
Google Translate	82.8	82.2	87.8	84.6	86.1	86.8	83.8	79.4	82.7	82.1	80.6
Gemini 2.5 Pro ⊕R	82.7	81.6	87.9	83.7	85.5	87.3	84.3	79.2	82.5	81.1	80.8
GPT-5 ⊕R	82.2	81.3	87.9	83.5	85.3	87.5	83.7	79.0	82.7	81.2	79.8
Claude 4.0 Sonnet ⊕R	83.2	83.4	87.1	83.3	85.5	86.6	83.9	78.2	82.8	81.5	80.4
Command A	81.3	81.8	87.4	82.7	86.2	87.6	82.4	75.9	82.6	81.2	78.5
DeepSeek R1	81.4	81.3	87.1	83.4	85.7	85.4	83.5	77.7	81.7	81.4	80.0
DeepL Pro	78.4	80.6	87.0	82.6	86.3	84.8	82.7	78.9	84.5	83.1	77.0
Qwen MT Plus	82.3	81.2	86.9	79.1	86.1	83.8	83.5	76.5	80.3	81.9	81.0
Mistral Medium 3.1	81.7	80.7	86.4	82.0	85.0	84.9	82.7	76.2	81.6	80.6	79.1
GPT-OSS 120B ⊕R	80.3	80.9	86.3	80.9	85.5	85.3	82.0	76.2	80.9	79.0	77.9
Llama 4 Maverick	79.5	78.9	86.1	81.4	85.2	85.3	82.2	75.3	80.9	79.1	78.4
Llama 3.1 405B	79.5	80.0	86.7	81.5	84.5	86.2	82.5	75.4	80.4	78.3	77.2
Qwen3-235B-A22B	79.4	80.7	85.6	80.7	85.5	85.2	82.1	74.5	80.3	81.2	79.4
Aya Expanse 32B	78.9	78.6	86.2	81.3	84.8	85.9	80.8	73.6	80.6	79.7	74.1
Gemma 3 (27b)	78.8	78.1	85.2	81.1	83.6	84.8	82.2	74.4	80.6	78.8	75.8
Mistral Large Latest	80.5	80.7	82.1	79.7	84.7	81.8	82.1	72.1	81.2	77.1	77.5

Table 7: Extended WMT24++ results with xComet-XL for extensive set of systems.

mark (ignoring white spaces and quotation marks). The final score is a percentage of failed cases. As the heuristic does not require reference translation, it can be easily scaled to any number languages. The only limitation is proper handling of question marks per language. We therefore also check for following language-specific question marks: Chinese and Arabic question mark, and the semi-colon for Greek.

We have not evaluated on Japanese, for which the question mark test doesn't work as the language allows different paraphrases not ending with question mark. An example from Google Translate: この文を翻訳せずに、「すべての星は星ですか？と言ってください」

GENDER1PERSON: Test Suite for estimating gender bias of first-person singular forms

Maja Popović[1], **Ekaterina Lapshinova-Koltunski**[2]
[1] IU University, Berlin, Germany
maja.popovic@iu.org
[2] University of Hildesheim, Germany
lapshinovakoltun@uni-hildesheim.de

Abstract

The GENDER1PERSON test suite is designed to measure gender bias in translating singular first-person forms from English into two Slavic languages, Russian and Serbian. The test suite consists of 1 000 Amazon product reviews, uniformly distributed over 10 different product categories. Bias is measured through a gender score ranging from -100 (all reviews are feminine) to 100 (all reviews are masculine).

The test suite shows that the majority of the systems participating in the WMT-2025 task for these two target languages prefer the masculine writer's gender. There is no single system which is biased towards the feminine variant. Furthermore, for each language pair, there are seven systems that are considered balanced, having the gender scores between -10 and 10.

Finally, the analysis of different products showed that the choice of the writer's gender depends to a large extent on the product. Moreover, it is demonstrated that even the systems with overall balanced scores are actually biased, but in different ways for different product categories.

1 Introduction

While English does not have many morphological forms related to gender, the two target languages do so. In those languages, gender marking exists not only for pronouns and animate nouns, but also for nouns, adjectives, verbs, determiners and numbers. If the text is written in the first-person singular form and no information about the gender of the author is provided, the translator can choose any of the two binary[1] genders. One of the most frequent affected POS tags are adjectives and past or passive participles. For example, "*I am happy that I bought this*" can be translated into Serbian in two ways, "*Srećan/srećna sam što sam ovo kupio/kupila*", depending on the writer's natural gender. This may result in translation errors, mismatches and inconsistencies, as well as in gender bias.

Our test suite is designed to measure bias of this type of gender in translations from English into Russian and Serbian. It consists of a carefully selected set of user reviews about Amazon products, because these texts are written in the first-person form and therefore very convenient. The test suite also enables the analysis of writer's gender depending on the product category. Although currently covering two target languages, it can easily be extended to more languages with similar rules for first-person singular gender.

Our main motivation was the results of our experiments reported in (Popovic and Lapshinova-Koltunski, 2024). We found some interesting tendencies regarding the writer's gender in user reviews of Amazon products from the DiHuTra corpus (Lapshinova-Koltunski et al., 2022). However, this corpus is designed for investigating differences between human and machine translations, but it is not tailored for exploring gender. The corpus is relatively small, only 196 reviews in total, and one third of them do not contain any indicator of the writer's gender. Therefore, the reported results (especially for different products), while interesting, were not fully reliable. The test suite presented in this paper, is much bigger and contains 1000 reviews.

2 Related work

While there is a large portion of work dealing with different types of gender bias, there are not many studies focussing on the first-person constructions. For example, Habash et al. (2019) propose automatic generation of both gender variants for the first person in Arabic NMT translations.

Our test suite also enables analysing bias depen-

[1] In the analysed target languages there are still no non-binary forms.

Proceedings of the Tenth Conference on Machine Translation, pages 800–822
November 8-9, 2025

dence on product category. Similarly, bias variation was addressed in (Zhao et al., 2017) who reported that data sets for specific tasks (e.g. cooking) contain significant gender bias and, furthermore, models trained on these data sets further amplify existing bias.

Our test suite uses a gender score as a metric. Cho et al. (2019) also proposes a measure of gender bias, however, in a completely different context: the metric measures the relation between gender and positive/negative expressions or occupations.

There exist other test suites. For instance, Stanovsky et al. (2019) design a test suite for evaluating gender bias in MT related to occupational nouns. Their method was developed for eight target languages, including Russian. Vanmassenhove and Monti (2021) create an English–Italian test suite with a focus on the resolution of natural gender. The authors provide word-level gender tags on the English source side and multiple alternative gender translations, where needed, on the Italian target side. Savoldi et al. (2023, 2024) present a test suite to investigate systems' ability to correctly translate the gender of the speaker int he context of occupations and professions (e.g. "*I am a doctor*"). The test suite from (Dawkins et al., 2024) measures the gender resolution tendencies of MT systems in literary-style dialogues.

However, to the best of our knowledge, none of the available test suites addresses the writer's gender from a general point of view. We believe that the presented test suite will add value to studies addressing gender in the area of machine translation and natural language processing.

3 Test suite creation

The test suite consists of 1 000 user reviews of 10 different product categories extracted from the publicly available repository "Amazon reviews 2023"[2] (Hou et al., 2024). The repository contains reviews written between 1996 and 2023 divided into more than 30 different product categories. For our test suite, we selected 10 categories, and extracted the first 100 reviews from each according to the following criteria:

- take the newest reviews, written in 2023;
- take the reviews with at least 10 occurrences of the first-person pronoun "I";
- take the reviews not longer than 250 (untokenised) words;
- do not include repeated reviews.

The threshold for the pronoun "I" is set to ensure that there will be enough instances of first-person singular gendered words in the translations. The length limit is set to avoid very long reviews, and very short reviews are automatically discarded due to the threshold for the pronoun "I".

The statistics of the obtained test suite is shown in Table 1. The product categories are selected to be distinct, and also to include some stereotypically masculine (e.g. cars, tools and improvements) and feminine (eg. beauty and baby products) as well as supposedly neutral ones (e.g. pet supplies).

In most of the reviews, the gender of the writer is not specified by any information in the English source. Explicit gender cues (e.g. "*I'm a man/woman*" or "*male/female*", "*I was pregnant*") can be found only in 1.5% of the reviews: there are 14 explicitly feminine reviews and one masculine. In addition, some of the reviews contain potential gender cues, namely "*husband*" and "*wife*" which used to be explicit and therefore can influence the choice of the writer's gender. In total, there are 17 reviews with "*my husband*" and 16 reviews with "*my wife*". All in all, there are notably more explicit feminine cues, while potential cues are balanced. Therefore, it can be expected that a translation with balanced gender distribution might contain slightly more feminine reviews. The distribution of the gender cues over product categories can be seen in Appendix A.1.

Validation set The validation set is created in order to check the performance of the evaluation scripts. The set is constructed according to the same rules as the test suite, the only differences are the included years and the size. The reviews were selected from all years before 2023, in order to avoid any overlap with the test suite. From each of the product categories, the first 10 reviews were selected so that it consists of 100 reviews in total, with 1 238 occurrences of the pronoun "I" and 18 789 untokenised running words.

The text is translated into Russian and Serbian using `Gemini 2.5 Flash` model[3] and `Google Translate`[4] in June 2025, and the four translations

[2] https://amazon-reviews-2023.github.io/

[3] https://gemini.google.com/u/1/app

[4] https://translate.google.com

product category	reviews	ocurrences of "I"	untokenised running words
AUTOMOTIVE	10x100	1 207	19 418
BABY PRODUCTS		1 222	19 051
BEAUTY AND PERSONAL CARE		1 229	19 054
HEALTH AND HOUSEHOLD		1 265	18 444
HOME AND KITCHEN		1 261	19 007
MUSICAL INSTRUMENTS		1 209	19 316
PET SUPPLIES		1 216	19 312
SPORTS AND OUTDOORS		1 218	18 562
TOOLS AND HOME IMPROVEMENT		1 207	19 410
VIDEO GAMES		1 286	18 345
total	1 000	12 240	189 919

Table 1: Corpus statistics: English Amazon reviews from ten different product categories: number of reviews, number of occurrences of the first-person pronoun "I", and length (number of untokenised running words).

are used for human assessment of the evaluation scripts described in the following sections.

4 Evaluation method

The evaluation method follows the principles from the manual gender labelling described in (Popovic and Lapshinova-Koltunski, 2024), but is fully automatic. It consists of three steps: (1) word-level annotation (identifying gendered words related to first-person singular), (2) review-level annotation based on the word-level gender labels, and (3) calculation of gender score based on the review-level gender labels.

4.1 Word-level annotation

The word-level annotation consists of identifying and labelling gendered words of interest, namely words referring to the first-person singular. For both languages, the words of interest are verb past participles and adjectives.

The annotation is based on POS tags from Stanza tool (Qi et al., 2020). For each language, a corresponding rule-based Python script is used. Two different scripts are necessary partly due to the differences between languages, and partly because of the differences between the provided POS tags.

Examples of tagged words for each of the languages can be seen in Table 2. For Serbian, both universal POS tags as well as treebank-specific POS tags which contain the information about person, gender, tense, number, etc. are available. For Russian, only universal POS tags are available and the further information can be found only in morpho-syntactic features.

From the given example, it can also be seen that, unfortunately, neither POS tags nor morpho-syntactic features of verb past participles and adjectives contain the information whether they correspond to the first-person singular. Therefore, a span of surrounding words has to also be checked according to the grammatical rules for each language.

In Serbian, words of interest can precede or follow the auxiliary verb "*biti*" (*to be*) in any first-person singular form. Since pronouns are in general more often omitted than not, they cannot be used here. Although the word order is rather free, the distance between the auxiliary verb and a word of interest is usually not larger than 3. Therefore, the context of 3 preceding and 3 following words was included. Increasing the range would serve only for a small number of cases, but would lead to picking up words referring to other persons or objects, thus decreasing the precision and potentially deteriorating the overall performance.

In Russian, words of interest follow the first-person singular personal pronoun я (*I*) and there is no auxiliary verb. In some cases, the pronoun can be placed immediately after the word of interest. Similarly to Serbian, while longer distances are also possible, increasing the span can easily decrease the precision. Therefore, the context of 3 preceding and one following word was included.

The overall process can be described as follows:

- find a potential word of interest (verb past participle or adjective);
- check whether the auxiliary verb/personal pro-

language	word	universal POS	treebank POS	universal morpho-syntactic features
Serbian	to	DET	Pd-nsn	Case=Nom|Gender=Neut|Number=Sing|PronType=Dem
	sam	AUX	**Var1s**	Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
	uradio	VERB	**Vmp-sm**	Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act
Russian	я	PRON	NA	**Case=Nom|Number=Sing|Person=1|PronType=Prs**
	этого	PRON	NA	Animacy=Inan|Case=Gen|Gender=Neut|Number=Sing|PronType=Dem
	делала	VERB	NA	Aspect=Imp|**Gender=Fem**|Mood=Ind|**Number=Sing**|**Tense=Past**|VerbForm=Fin|Voice=Act

Table 2: Examples of words annotated by Stanza tool in Serbian (above) and Russian (below).

noun in first-person singular form can be found in the given context;

- if yes, take the gender of the word of interest and increment the corresponding gender count.

The main difference between the two scripts is related to morpho-syntactic features. While they were immediately available in Serbian tree-bank tags, finding them in Russian required more computational effort. First, the universal POS tag was checked, and then the list of morpho-syntactic features was traversed in order to find additional information about the gender, number, as well as person of surrounding words. For these reasons, the script for Russian is much slower than the one for Serbian.

4.2 Human assessment on the validation set

In order to assess the performance of the word-level annotation, the two `Gemini` translations of the validation set described in Section 3 were annotated by the corresponding scripts. The annotations are then checked by experts, i.e. trained linguists with the native command of the target languages. They were instructed to determine whether the annotated words are correct (precision) as well as whether all gendered words of interest were captured (recall). The results in Table 3 show very high precision (over 99%) for both languages, meaning that almost all annotated words are really referring to the first-person singular. As for recall, it is high for Serbian (95%), but notably smaller for Russian (75%).

A qualitative analysis of errors showed that in both languages, the long-range dependencies were not captured, as expected. Further analysis of the low Russian recall revealed that for a considerable number of frequent adjectives, the relevant information about the nominative case was missing, so that they were not considered as words of interest. If the rule were changed, many other adjectives (referring to other people or objects) would be selected thus decreasing the precision and possibly deteriorating the overall performance.

script validation		
	en-ru	en-sr
precision	99.8	99.7
recall	75.6	95.2

Table 3: Evaluation of annotating scripts on a validation set.

Another problem with Russian is occasional occurrence of informal style where the pronoun is fully omitted. For those cases, it is practically impossible to create a rule for capturing the word of interest because there are no related first-person singular words around.

Because of the low recall for Russian, the review-level annotation (described in the next section) which is essential for the task was manually checked as well, on all four translations of validation set. The resulting scores can be seen in Appendix A.2. Since the review labels were correct and no problems related to word annotation errors were identified, the script is considered well-suited for the task.

It should be noted that the problems with low word-level recall could be addressed by using LLMs. Our initial experiment with LLMs using few-shot prompts was, however, not successful. While being able to increase the recall to some

extent, the precision dropped notably because of tagging a large number of irrelevant words (second and third person referring to other people or objects, often even in plural form). A systematic set of experiments with different prompt designs would be necessary, and will be investigated in future work.

4.3 Review-level annotation

For each review, a gender label is assigned according to the gender of the identified words of interest in the previous step. If no words of interest were identified, the review is labelled as "x" (no gender found). Otherwise, a gender proportion score

$$gp = \frac{C(m) - C(f)}{C(m) + C(f)}$$

is calculated, where $C(f)$ denotes the count of feminine words of interest and $C(m)$ presents the masculine count.

$$gp = \begin{cases} \text{feminine}, & gp < -0.4 \\ \text{masculine}, & gp > 0.4 \\ \text{mixed}, & -0.4 \leq gp \leq 0.4 \end{cases}$$

If the $gp < -0.4$, the review is considered feminine, if $gp > 0.4$ masculine. If the score is between -0.4 and 0.4, the review is considered as "mixed". This soft decision approach is chosen to alleviate potential errors of the annotation scripts and also to retain clear tendencies of a translation model towards one gender on the word level.

Table 4 presents an example from the validation set. In the Russian translation, all words of interest are masculine, so that the gender proportion gp is equal to 1 and the review-level gender is masculine. In the Serbian translation, there are five feminine and two masculine words. The gender proportion is then $gp = (2-5)/(2+5) = -3/7 = -0.43$. Since it is less than -0.4, the review is labelled as feminine. If there were four feminine and two masculine words, the proportion would be $(2-4)/(2+4) = -2/6 = -.033$, which is between -0.4 and 0.4 so that the review would be labelled as mixed.

4.4 Gender scores

In order to estimate gender bias in a set of reviews, the following score is calculated:

$$genderScore = 100 * \frac{N(m) - N(f)}{N}$$

where $N(m)$ denotes the total number of masculine reviews, $N(f)$ is the total number of feminine reviews, and N is the total number of reviews, including those marked with "x" and the mixed ones. The "x" and mixed reviews are thus not contributing to the score.

It should be explained that the analysis of the "x" reviews goes beyond the scope of this work, so it is not known whether they are really genderless (not containing any words of interest), or written in the gender-neutral way (difficult for both target languages but possible in some cases), or written using some kind of inclusive forms. However, there are not so many systematic consistent strategies for gender-neutral or inclusive forms. They may include the use of plural verb forms with first person singular pronouns or gender-gapping (the use of underscore, e.g. студент_ка (student(m/f)) and also the use of impersonal or indefinite personal structures. It is also common to use both forms, e.g. я был/а разочарован/а or *bio/la sam razočaran/a* (I was(m/f) disappointed(m/f)) in Russian and Serbian, respectively[5]. The POS tagger cannot properly recognise these inclusive forms, and would label them as (proper) nouns.

The values of the gender score range from -100 (all reviews in the text are feminine) to 100 (all reviews are masculine). There are no "good" and "bad" values as such, only the information about the (dis)balance of the two genders in a text. Negative values indicate more feminine reviews, positive values indicate more masculine reviews, and the smaller the absolute value is, the more gender-balanced the text is. We consider the texts with a score between -10 and 10 as gender-balanced.

5 Results on the WMT-2025 translations

5.1 Evaluation levels

In the framework of WMT-2025, the test suite was translated by 40 English→Russian systems and 35 English→Serbian systems. For each language pair, the gender scores are calculated in the following set-ups:

1. language level (all systems together);
2. system level;
3. language level for each product category;
4. system level for each product category.

[5] See more details in (Popovic and Lapshinova-Koltunski, 2024; Popović et al., 2025).

	en	I **came** across an item online with the same concept and **was** completely **interested**. I eventually **bought** them. They were too big for what I **wanted**. I then **bought** these. I **was thinking** of purchasing more.
masc.	ru	Я наткнулся в интернете на предмет с такой же концепцией и полностью заинтересовался. В конце концов я купил их. Они были слишком большими для того, что я хотел. Затем я купил эти. Я думал о покупке еще.
fem.	sr	Naišla sam na predmet na internetu sa istim konceptom i bio sam potpuno zainteresovan. Na kraju sam ih kupila. Bili su preveliki za ono što sam želela. Onda sam kupila ove. Razmišljala sam o kupovini više.

Table 4: Example of gender labels according to first-person singular gendered words.

overall results for each language pair					
		distribution			
lang.	score	m	f	x	mix
en-ru	33.8	25 934	12 427	654	985
en-sr	39.0	23 555	9 920	159	1 366

Table 5: Language level gender scores and label distributions

The gender scores are presented together with the distribution of the review-level gender labels which led to the score value. In the following sections, the results for the first three set-ups are presented and discussed in detail, and the results for the fourth set-up are presented and discussed only for gender-balanced systems. The complete results for all systems and all product categories together with the corresponding discussions are presented in Appendix A.3.

5.2 Language level scores

Overall gender scores for each of the target languages aggregated over all translation outputs are presented in Table 5.

The gender scores are between 30 and 40, indicating that for both languages the majority of the reviews are translated as masculine.

Looking at the distributions of review-level labels, it can be noted that the counts of masculine reviews are similar in both languages, and notably higher than the counts of feminine labels. Furthermore, the number of mixed reviews is notably higher in Serbian, while the number of "x" labels is significantly higher (about 4 times) in Russian. A possible reason for more mixed labels in Serbian is that Serbian is less-resourced than Russian so that there are more translation errors. As for the larger amount of "x" labels in Russian, one possible reason is that Russian systems often generate some kind of inclusive form. Another possibility is the low recall of the annotation script, so that in some of the reviews none of the words of interest are captured. As previously mentioned, the nature of "x" labels was not analysed in this work, but should be part of the future research.

5.3 System level scores

The gender scores and label distributions for each of the participating systems are presented in Tables 6 and 7. The systems are ranked from lowest to highest scores, and, as previously mentioned, the most balanced systems are considered to be those with scores between -10 and 10. It can be noted that for both languages, only seven systems have balanced score values. Five of them, namely `Algharb`, `Gemini-2.5 Pro`, `Shy`, `Wenuiil` and `Yolu` are balanced for both languages. `GemTrans` is among the most balanced for Serbian, but also not very far for Russian with the score of 14.9. `ONLINE-G` and `ONLINE-B` have different tendencies in the two languages: balanced for one, but very (46.9) or extremely (86.7) masculine for the other. Moreover, it can be noted that `ONLINE-G` (in contrast to other balanced systems) generated a high number of mixed genders in both languages. `Yandex` did not participate in translating into Serbian.

Furthermore, it can be seen that all other systems are masculine-biased, to more or less extent. There is no system with a bias towards feminine writer's gender. Moreover, two systems, `TranssionMT` and `Mistral-7B`, are extremely biased for both languages, with (almost) all reviews translated into masculine form – not a single feminine review was identified in `TranssionMT` outputs.

5.4 Language level scores for different product categories

Table 8 presents the language level gender scores for each of the ten product categories. The tenden-

English→Russian					
		distribution			
system	score	m	f	x	mix
Algharb	-0.5	485	490	16	9
Yolu	-1.2	482	494	13	11
Yandex	-1.8	481	499	17	3
Gemini-2.5-Pro	1.7	499	482	16	3
ONLINE-G	2.3	412	389	10	189
Wenyiil	8.3	528	445	19	8
Shy	9.9	533	434	27	6
Laniqo	11.6	537	421	31	11
GemTrans	14.9	568	419	9	4
SalamandraTA	16.8	561	393	15	31
TowerPlus-9B	17.1	579	408	9	4
IRB-MT	18.3	580	397	18	5
hybrid	20.3	580	377	34	9
Claude-4	20.4	590	386	20	4
Gemma-3-12B	23.1	603	372	15	10
GPT-4.1	23.3	607	374	18	1
DeepSeek-V3	24.4	607	363	26	4
ONLINE-W	25.5	528	273	11	188
DLUT_GTCOM	26.5	614	349	20	17
UvA-MT	27.4	630	356	10	4
AyaExpanse-32B	31.7	649	332	11	8
Qwen3-235B	34.4	662	318	12	8
EuroLLM-22B	35.5	665	310	12	13
Gemma-3-27B	38.2	681	299	13	7
CommandA	39.3	686	293	14	7
TowerPlus-72B	40.6	693	287	9	11
TranssionTranslate	44.7	645	198	10	147
ONLINE-B	46.9	715	246	10	29
AyaExpanse-8B	47.6	728	252	9	11
IR-MultiagentMT	47.6	719	243	30	8
Qwen2.5-7B	48.0	681	201	60	58
SRPOL	49.2	733	241	12	14
Llama-4-Maverick	54.0	762	222	13	3
CommandA-MT	54.8	767	219	8	6
CommandR7B	55.0	753	203	17	27
Llama-3.1-8B	55.2	750	198	5	47
EuroLLM-9B	66.7	822	155	17	6
NLLB	83.6	896	60	29	15
Mistral-7B	90.9	938	29	3	30
TranssionMT	98.5	985	0	6	9

Table 6: System level gender scores and review label distributions for Russian

English→Serbian					
		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	-3.2	483	515	1	1
Algharb	-2.3	485	508	1	6
ONLINE-B	3.7	506	469	1	24
Yolu	4.4	510	466	2	22
GemTrans	6.8	527	459	2	12
Wenyiil	7.0	531	461	1	7
Shy	7.6	535	459	1	5
CUNI-SFT	16.0	549	389	5	57
GPT-4.1	20.7	601	394	1	4
Claude-4	21.4	604	390	1	5
EuroLLM-22B	22.9	589	360	2	49
hybrid	22.9	609	380	3	8
IRB-MT	23.3	608	375	2	15
Gemma-3-12B	27.9	606	327	2	65
AyaExpanse-32B	29.9	622	323	3	52
Gemma-3-27B	32.4	645	321	3	31
UvA-MT	32.5	656	331	1	12
DeepSeek-V3	34.5	668	323	1	8
TowerPlus-9B	36.4	628	264	22	86
AyaExpanse-8B	37.8	626	248	5	121
Qwen3-235B	43.8	715	277	1	7
CommandR7B	45.1	643	192	54	111
Llama-3.1-8B	46.7	707	240	3	50
CommandA	52.4	747	223	1	29
IR-MultiagentMT	52.6	753	227	6	14
EuroLLM-9B	56.9	750	181	2	67
Qwen2.5-7B	58.7	732	145	12	111
SalamandraTA	59.9	765	166	3	66
Llama-4-Maverick	62.5	804	179	2	15
CommandA-MT	69.6	841	145	1	13
TowerPlus-72B	70.8	835	127	3	35
Mistral-7B	86.1	898	37	4	61
ONLINE-G	86.7	878	11	3	108
TranssionMT	90.8	916	8	2	74
TranssionTranslate	98.3	983	0	2	15

Table 7: System level gender scores and review label distributions for Serbian

cies are the same in both languages: feminine bias is present only for two product categories: BEAUTY AND PERSONAL CARE and BABY PRODUCTS. And even though they are clearly "feminine", the gender scores are not lower than -45.

For all other products categories, the majority of the reviews are masculine. The most balanced category is PET SUPPLIES, with the score of 5.5 for Russian and 19.0 for Serbian.

The most masculine products seem to be VIDEO GAMES, MUSICAL INSTRUMENTS and AUTOMOTIVE, with all gender scores over 80. While the biases in BEAUTY AND PERSONAL CARE and BABY PRODUCTS as well as in VIDEO GAMES and AUTOMOTVE are expected due to the widely known stereotypes, the results for MUSICAL INSTRUMENTS are somewhat surprising. The same tendency was already observed in (Popovic and Lapshinova-Koltunski, 2024). However, the results were reported only on a small data set consisting of 14 reviews per category, so they were not reliable. It was nevertheless striking that even the human translators did not opt for the feminine writer's gender for any of the reviews in this category.

As for the rest of the categories, the most masculine one is TOOLS AND HOME IMPROVEMENT with the scores over 60, followed by SPORTS AND OUTDOORS with the scores around 45. The other two, HEALTH AND HOUSEHOLD and HOME AND KITCHEN are more balanced, with scores between 20 and 30.

The label "x" is relatively uniformly distributed over the categories, except BEAUTY AND PERSONAL CARE in Russian, and BABY PRODUCTS in both languages, where a notably higher amount of "x" reviews can be noted. A deeper linguistic analysis for a better understanding of the nature of these translations could, therefore, start from translations of reviews in these product categories.

The amount of mixed reviews is apparently proportional to the amount of feminine reviews. The tendency is confirmed by calculating Pearson correlation coefficients presented in Table 9 (low correlations for the "x" label can be seen, too).

A possible reason might be that the models are intrinsically inclined to choose masculine first-person singular words, so that even when the number of feminine words increases, many of them are still mixed with masculine words within the same review. A deeper analysis of the word-level annotations might reveal more details about the background, and is planned for the future work.

5.5 System level scores for different product categories

The gender scores of each product category for the gender-balanced systems are presented in Table 10. It can be seen that, although the systems are gender-balanced for the entire test suite, they are far from balanced within different product categories. This means that the reason for the overall balance lies in the choice of the product categories and not in the properties of the systems. They are all heavily masculine-biased for each of the three most masculine categories, namely AUTOMOTIVE, MUSICAL INSTRUMENTS and VIDEO GAMES. The overall balance seems to be achieved because this masculine bias is compensated not only by the two most feminine categories BABY PRODUCTS and BEAUTY AND PERSONAL CARE, but also by HEALTH AND HOUSEHOLD, HOME AND KITCHEN and PET SUPPLIES.

As for differences between the languages, ONLINE-G is balanced for Russian but clearly masculine-biased for Serbian for all categories. ONLINE-B is balanced for Serbian, whereas for Russian, HEALTH AND HOUSEHOLD, HOME AND KITCHEN and PET SUPPLIES are predominantly masculine instead of feminine, and the feminine bias for Baby Products is notably smaller.

Other system behave differently for different product categories and no regular patterns were observed, although there are certain tendencies which are discussed in Appendix A.3.

6 Summary and Outlook

Summary The presented test suite is designed for analysis of the first-person singular gender (speaker's or writer's gender) in translations from English into Russian and Serbian. The gender score is defined to measure the balance between the two binary genders, masculine and feminine. The score ranges between -100 (fully feminine) and 100 (fully masculine), and values between -10 and 10 are considered as balanced.

After using the test suite on WMT-2025 translation outputs to calculate language level scores, system level scores and product level scores, the main findings are:

- the majority of the systems are biased towards masculine writer's gender;
- none of the systems is biased towards feminine writer's gender;

product category	lang.	score	distribution m	f	x	mix
AUTOMOTIVE	en-ru	83.0	3610	291	50	49
	en-sr	81.5	3141	287	14	58
BABY PRODUCTS	en-ru	-28.1	1324	2447	89	140
	en-sr	-14.1	1380	1875	49	196
BEAUTY AND PERSONAL CARE	en-ru	-42.7	1033	2742	116	109
	en-sr	-27.6	1155	2122	16	207
HEALTH AND HOUSEHOLD	en-ru	21.9	2338	1463	67	132
	en-sr	26.9	2130	1189	10	171
HOME AND KITCHEN	en-ru	19.7	2299	1512	58	131
	en-sr	30.3	2192	1132	12	164
MUSICAL INSTRUMENTS	en-ru	84.5	3648	269	47	36
	en-sr	82.5	3157	270	12	61
PET SUPPLIES	en-ru	5.5	2000	1780	49	171
	en-sr	19.0	1986	1322	13	179
SPORTS AND OUTDOORS	en-ru	45.6	2836	1011	43	110
	en-sr	46.7	2494	860	13	133
TOOLS AND HOME IMPROVEMENT	en-ru	63.2	3191	662	68	79
	en-sr	60.9	2755	625	8	112
VIDEO GAMES	en-ru	85.1	3655	250	67	28
	en-sr	83.6	3165	238	12	85

Table 8: Results for each product category and each language pair

	m	f
mix	-.839	.714
x	-.238	.497

Table 9: Pearson's correlation coefficients between the gender labels in different product categories: the number of mixed reviews is proportional to the number of feminine reviews.

- five systems are gender-balanced for both target languages (with scores between -10 and 10): *Algharb*, *Gemini-2.5-Pro*, *Shy*, *Wenyiil* and *Yolu*;
- one model (*GemTrans*) is only slightly unbalanced towards masculine for Russian (score 14.9)
- two models behave differently depending on the language (*ONLINE-B* and *ONLINE-G*)
- one model (*Yandex*) participated only for Russian

Further analysis of different product categories showed that none of the systems is balanced within all product categories, while some systems are balanced within one single product category. This means that the overall gender balance is a consequence of the choice of product categories, not of the system properties.

Furthermore, three product categories are identified to be predominantly masculine and two as mostly feminine. AUTOMOTIVE, MUSICAL INSTRUMENTS and VIDEO GAMES are heavily biased towards masculine by all models (all scores are larger than 50). BABY PRODUCTS and BEAUTY AND PERSONAL CARE are biased towards feminine, but to a lesser extent: scores are ranging from -80 to 99, and a few systems are balanced.

Overall, the results confirmed the similar tendencies reported in previous work for Croatian and Russian on a small scale (Popovic and Lapshinova-Koltunski, 2024). Given that one of the reported tendencies was that even human translators are generally inclined to opt more often for a masculine writer, and are also influenced by the product category, the WMT-2025 results from the test suite are not surprising.

Outlook The test suite offers several possibilities for future work, some of them are already mentioned in previous sections. One important open

system		all	Auto	Baby	Beauty	Health	Home	Music	Pets	Sports	Tools	Games
Algharb	ru	-0.5	75	-79	-74	-36	-29	74	-61	**8**	38	79
	sr	-2.3	72	-83	-79	-32	-35	77	-61	**6**	34	78
Gemini	ru	1.7	71	-77	-73	-31	-22	74	-52	17	33	77
	sr	-3.2	74	-83	-77	-36	-32	72	-66	**4**	32	80
GemTrans	ru	*14.9*	85	-64	-74	**-9**	-17	82	-37	35	58	90
	sr	6.8	74	-60	-81	-30	-17	74	-39	28	38	81
ONLINE-B	ru	*46.9*	*93*	*-24*	*-62*	*52*	*40*	*95*	*38*	*62*	*77*	*98*
	sr	3.7	77	-72	-84	-43	-26	80	-47	31	33	88
ONLINE-G	ru	2.3	58	-43	-79	-53	-49	76	-29	24	39	79
	sr	*86.7*	*97*	*63*	*74*	*88*	*82*	*98*	*89*	*89*	*91*	*96*
Shy	ru	9.9	76	-65	-69	-15	**-10**	76	-41	22	50	75
	sr	7.6	72	-73	-74	-23	-12	78	-43	27	46	78
Wenyiil	ru	8.3	81	-75	-71	-20	-16	75	-39	17	48	83
	sr	7.0	74	-74	-74	-15	-12	86	-47	**10**	44	78
Yolu	ru	-1.2	71	-67	-73	-30	-45	78	-50	12	29	63
	sr	4.4	72	-65	-86	-22	-26	70	-39	22	41	77
Yandex	ru	-1.8	73	-73	-75	-39	-31	72	-63	22	36	60

Table 10: Gender scores for the most balanced systems: overall, and for each product category. Balanced product category scores are presented in bold.

question is the nature of the translations with the label "x". It should be explored whether they are really gender-neutral, or an inclusive form was used, or the annotation script missed all words of interest, or there are possibly translation errors. As for mixed reviews, word-level gender scores could be added to include their contribution. Another direction is analysis of reviews with gender cues (for instance specific words like *pregnant*, etc., and their possible separation into a sub-suite.

An obvious direction is extending the test suite with more data. Also, it can be easily extended to other languages with gendered first-person singular words, such as French, Spanish, Czech among others. Also, other domains/genres apart from Amazon product reviews should be considered.

Finally, a systematic experiment on prompt design should be carried out in order to use LLMs for word-level annotation and improve recall without deteriorating precision.

Limitations

The presented test suite comes with a few limitations. Currently, it deals only with two target languages, both of them being Slavic although with different grammar rules. Furthermore, only one evaluation method has been systematically explored so far, namely using Stanza POS tags for a rule-based identification of gendered words. The script performs well for Serbian, but has notably lower recall for Russian, and also a high time complexity, due to the differences between the POS tags provided by the tool. Moreover, it should be kept in mind that mixed and "x" reviews were excluded from the gender score, but not analysed. Also, the reviews with explicit or implicit gender cues are not analysed.

References

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Hillary Dawkins, Isar Nejadgholi, and Chi-kiu Lo. 2024. WMT24 test suite: Gender resolution in speaker-listener dialogue roles. In *Proceedings of the Ninth Conference on Machine Translation*, pages 307–326, Miami, Florida, USA. Association for Computational Linguistics.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language

and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Ekaterina Lapshinova-Koltunski, Maja Popović, and Maarit Koponen. 2022. DiHuTra: a parallel corpus to analyse differences between human translations. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1751–1760, Marseille, France. European Language Resources Association.

Maja Popovic and Ekaterina Lapshinova-Koltunski. 2024. Gender and bias in Amazon review translations: by humans, MT systems and ChatGPT. In *Proceedings of the 2nd International Workshop on Gender-Inclusive Translation Technologies*, pages 22–30, Sheffield, United Kingdom. European Association for Machine Translation (EAMT).

Maja Popović, Ekaterina Lapshinova-Koltunski, and Anastasiia Göldner. 2025. Did I (she) or I (he) buy this? or rather I (she/he)? towards first-person gender neutral translation by LLMs. In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 64–73, Geneva, Switzerland. European Association for Machine Translation.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in mt with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation*, pages 252–262, Singapore. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. FBK@IWSLT test suites task: Gender bias evaluation with MuST-SHE. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 65–71, Bangkok, Thailand (in-person and online). Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Eva Vanmassenhove and Johanna Monti. 2021. gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

A Appendix

A.1 Explicit and potential gender cues

Table 11 presents the distribution of explicit ("*I am a man/woman*" or "*male/female*", "*I am pregnant*", etc.) and potential ("*my husband*", "*my wife*") gender cues over different product categories, mentioned in Section 3.

The single explicit masculine cue occurs in the AUTOMOTIVE category. The explicit feminine cues are distributed over several categories, most frequently in BABY PRODUCTS and TOOLS AND HOME IMPROVEMENT followed by BEAUTY AND PERSONAL CARE and SPORTS AND OUTDOORS.

A.2 Review labels on the validation set

Table 12 presents gender scores and label distributions for two translations of the validation set described in Section 3. The evaluation scripts were run on the two translations of the validation set, the one generated by Gemini (which was also used for assessing word-level annotation), and another one generated by Google Translate. The review labels were checked by human annotators in order to assess the influence of word-level errors on the review-level labels (and thus on the final score).

A.3 Product categories: system level

This section presents and discusses the system level scores for each product category (Tables 13–32), as mentioned in Section 5. While systems generally behave differently for different products, certain tendencies can be observed in each of the product categories. The overall gender balanced systems are presented in italic. It is already discussed in Section5 that they are far from balanced for most of the categories, being clearly biased towards one or other writer's gender. In this section it can be seen that they are less biased in the three heavily masculine categories than other systems, and more biased in the two heavily feminine categories than other systems, but also clearly feminine biased in categories with relatively uniform distribution of feminine and masculine scores (e.g. HEALTH AND HOUSEHOLD, HOME AND KITCHEN and PET SUPPLIES.

	feminine cues		masculine cues	
	explicit	potential	explicit	potential
AUTOMOTIVE	0	1	1	3
BABY PRODUCTS	3	1	0	0
BEAUTY AND PERSONAL CARE	2	2	0	0
HEALTH AND HOUSEHOLD	1	0	0	2
HOME AND KITCHEN	1	4	0	0
MUSICAL INSTRUMENTS	1	2	0	4
PET SUPPLIES	0	2	0	0
SPORTS AND OUTDOORS	2	4	0	2
TOOLS AND HOME IMPROVEMENT	3	0	0	1
VIDEO GAMES	0	1	0	3
total	14	17	1	16

Table 11: Number of reviews (in each product category and overall) with explicit (man, woman, pregnancy, female, male) and potential (husband, wife) gender cues.

			distribution			
	system	score	m	f	x	mix
en-ru	Gemini	-17.1	40	57	3	0
	Google	48.0	71	23	3	3
en-sr	Gemini	-20.0	39	59	1	1
	Google	4.0	48	44	3	5

Table 12: Gender scores and label distributions for two translations of the validation set

AUTOMOTIVE (Tables 13 and 14) This category is overall heavily masculine-biased, with all system scores over 50. Even the overall gender balanced systems are clearly masculine for these products, with scores between 70 and 80.

BABY PRODUCTS (Tables 15 and 16) The majority of gender scores is feminine-biased, however, there are a few notably masculine-biased outputs. There are 6 balanced systems for Russian and 4 for Serbian, and none of them is balanced for both languages. As for overall gender balanced systems, all of them are clearly feminine-biased with the scores ranging from -83 to -43.

BEAUTY AND PERSONAL CARE (Tables 17 and 18) The majority of gender scores is feminine-biased, however, there are a few notably masculine-biased outputs. There are 2 balanced systems for Russian and 5 for Serbian. `CommandR7B` is balanced for both languages, more inclined to feminine for Russian and to masculine in Serbian. Interestingly, `Llama-3.1-8B` for Serbian is perfectly gender-balanced with the score equal to 0, though with 9 mixed and one "x" review. All overall gender balanced systems are clearly feminine-biased with the scores ranging from -86 to -69.

HEALTH AND HOUSEHOLD (Tables 19 and 20) There are both feminine and masculine gender scores, however more systems are masculine-biased. There are 6 balanced systems for Russian and 2 for Serbian. `Gemma-3-12B` is balanced for both languages, more inclined to feminine for Russian and to masculine in Serbian. All overall gender balanced systems are feminine-biased, although to a less extent than for the previous two categories, with the scores ranging from -53 to -12.

HOME AND KITCHEN (Tables 21 and 22) There are both feminine and masculine gender scores, however more systems are masculine-biased. There are 10 balanced systems for Russian and 3 for Serbian. *Claude-4* and `GPT-4.1` are balanced for both languages. `GPT-4.1` is more inclined to masculine for Russian and to feminine in Serbian. `Claude-4` is inclined to feminine in Russian and perfectly balanced in Serbian, with the score of 0 and no mixed or "x" reviews. Another perfectly balanced system is `DLUT_GTCOM` for Russian, however with 2 mixed and 2 "x" reviews. All overall gender balanced systems are feminine-biased, although to a less extent than for the previous three categories, with the scores ranging from -49 to -10 (`Shy` for Russian is considered as balanced with the score -10 exactly on the threshold).

MUSICAL INSTRUMENTS (Tables 23 and 24) This category is overall heavily masculine-biased, with all system scores over 50. Even the overall gender-balanced systems are heavily masculine-

biased with the scores between 70 and 86.

PET SUPPLIES (Tables 25 and 26) There are both feminine and masculine gender scores, relatively balanced proportion of systems in Russian however more masculine systems in Serbian. There are 6 balanced systems for Russian and 3 for Serbian. `AyaExpanse-32B` is balanced for both languages, more inclined to feminine for Russian and to masculine in Serbian. All overall gender balanced systems are feminine-biased, with the scores ranging from -66 to -29.

SPORTS AND OUTDOORS (Tables 27 and 28) There are no feminine gender scores, and there are two balanced systems (inclined to masculine) for each language. For Serbian, both those systems, `Algharb` and `Gemini-2.5-Pro`, are also balanced overall, while for Russian it is only `Algharb`. Other overall balanced systems are clearly although not heavily masculine, with the scores ranging from 12 to 31.

TOOLS AND HOME IMPROVEMENT (Tables 29 and 30) There are no feminine gender scores, and no balanced systems: the scores range from 29 to 100. Although this category is not so heavily masculine-biased as others, even the overall gender balanced systems are masculine-biased for this one with the scores between 29 and 50.

VIDEO GAMES (Tables 31 and 32) This category is overall heavily masculine-biased, with all system scores over 50. Even the overall gender-balanced systems are very masculine-biased with the scores ranging between 60 and 88.

AUTOMOTIVE, English→Russian					
		distribution			
system	score	m	f	x	mix
SalamandraTA	57.0	77	20	1	2
ONLINE-G	*58.0*	69	11	0	20
Laniqo	64.0	82	18	0	0
Gemini-2.5-Pro	*71.0*	85	14	1	0
Yolu	*71.0*	85	14	0	1
Yandex	*73.0*	86	13	1	0
TowerPlus-9B	74.0	86	12	1	1
Algharb	*75.0*	86	11	1	2
CommandR7B	76.0	86	10	2	2
Shy	*76.0*	87	11	2	0
UvA-MT	81.0	90	9	1	0
Wenyiil	*81.0*	90	9	1	0
DeepSeek-V3	82.0	90	8	2	0
IRB-MT	82.0	89	7	4	0
Qwen3-235B	82.0	90	8	1	1
AyaExpanse-32B	83.0	91	8	0	1
Gemma-3-12B	83.0	90	7	2	1
Qwen2.5-7B	83.0	88	5	5	2
Claude-4	84.0	91	7	2	0
AyaExpanse-8B	85.0	92	7	1	0
GemTrans	85.0	92	7	1	0
SRPOL	85.0	92	7	1	0
GPT-4.1	86.0	92	6	2	0
Llama-3.1-8B	86.0	92	6	1	1
TranssionTranslate	86.0	89	3	1	7
TowerPlus-72B	87.0	93	6	1	0
DLUT_GTCOM	88.0	93	5	2	0
hybrid	88.0	93	5	2	0
CommandA-MT	89.0	94	5	1	0
Gemma-3-27B	89.0	94	5	1	0
ONLINE-W	89.0	92	3	0	5
EuroLLM-22B	90.0	95	5	0	0
Llama-4-Maverick	90.0	94	4	2	0
IR-MultiagentMT	91.0	94	3	2	1
EuroLLM-9B	92.0	95	3	2	0
CommandA	93.0	96	3	1	0
ONLINE-B	93.0	96	3	1	0
NLLB	95.0	97	2	1	0
Mistral-7B	97.0	98	1	0	1
TranssionMT	99.0	99	0	0	1

Table 13: AUTOMOTIVE, Russian

AUTOMOTIVE, English→Serbian					
		distribution			
system	score	m	f	x	mix
CUNI-SFT	59.0	79	20	1	0
AyaExpanse-32B	66.0	82	16	0	2
CommandR7B	70.0	80	10	6	4
EuroLLM-22B	70.0	83	13	0	4
Algharb	*72.0*	86	14	0	0
Shy	*72.0*	86	14	0	0
Yolu	*72.0*	84	12	1	3
Gemini-2.5-Pro	*74.0*	87	13	0	0
GemTrans	*74.0*	87	13	0	0
UvA-MT	74.0	87	13	0	0
Wenyiil	*74.0*	87	13	0	0
ONLINE-B	*77.0*	88	11	0	1
AyaExpanse-8B	78.0	86	8	0	6
Claude-4	78.0	89	11	0	0
EuroLLM-9B	78.0	87	9	0	4
IRB-MT	78.0	89	11	0	0
Gemma-3-12B	79.0	87	8	0	5
TowerPlus-9B	80.0	87	7	1	5
Qwen2.5-7B	81.0	86	5	2	7
hybrid	82.0	91	9	0	0
GPT-4.1	84.0	92	8	0	0
Llama-3.1-8B	86.0	92	6	0	2
DeepSeek-V3	87.0	93	6	0	1
Gemma-3-27B	87.0	93	6	0	1
SalamandraTA	87.0	92	5	0	3
Qwen3-235B	88.0	94	6	0	0
IR-MultiagentMT	89.0	93	4	2	1
CommandA	90.0	95	5	0	0
TowerPlus-72B	91.0	95	4	0	1
CommandA-MT	92.0	96	4	0	0
Mistral-7B	94.0	96	2	1	1
ONLINE-G	97.0	97	0	0	3
TranssionMT	97.0	97	0	0	3
Llama-4-Maverick	98.0	99	1	0	0
TranssionTranslate	99.0	99	0	0	1

Table 14: AUTOMOTIVE, Serbian

BABY PRODUCTS, English→Russian					
		distribution			
system	score	m	f	x	mix
Algharb	*-79.0*	9	88	2	1
Gemini-2.5-Pro	*-77.0*	10	87	2	1
Wenyiil	*-75.0*	10	85	3	2
Yandex	*-73.0*	12	85	3	0
Yolu	*-67.0*	15	82	2	1
Shy	*-65.0*	16	81	3	0
GemTrans	-64.0	17	81	2	0
hybrid	-62.0	17	79	3	1
Claude-4	-58.0	20	78	1	1
GPT-4.1	-58.0	19	77	3	1
IRB-MT	-56.0	21	77	1	1
DeepSeek-V3	-54.0	21	75	3	1
Laniqo	-51.0	22	73	4	1
Gemma-3-12B	-47.0	25	72	2	1
UvA-MT	-46.0	26	72	1	1
DLUT_GTCOM	-44.0	24	68	3	5
ONLINE-G	-43.0	17	60	1	22
Gemma-3-27B	-40.0	28	68	3	1
SalamandraTA	-39.0	28	67	3	2
AyaExpanse-32B	-38.0	29	67	4	0
CommandA	-36.0	30	66	3	1
TowerPlus-9B	-36.0	31	67	2	0
ONLINE-W	-34.0	17	51	2	30
Qwen3-235B	-30.0	33	63	3	1
EuroLLM-22B	-29.0	34	63	2	1
ONLINE-B	-24.0	36	60	1	3
TowerPlus-72B	-23.0	36	59	3	2
SRPOL	-17.0	39	56	2	3
Qwen2.5-7B	-14.0	38	52	3	7
IR-MultiagentMT	-10.0	43	53	3	1
TranssionTranslate	-10.0	32	42	1	25
CommandA-MT	-6.0	46	52	1	1
AyaExpanse-8B	-4.0	46	50	2	2
Llama-4-Maverick	-4.0	47	51	2	0
Llama-3.1-8B	3.0	48	45	1	6
CommandR7B	14.0	54	40	2	4
EuroLLM-9B	22.0	59	37	3	1
NLLB	74.0	85	11	2	2
Mistral-7B	80.0	87	7	0	6
TranssionMT	97.0	97	0	2	1

Table 15: BABY PRODUCTS, Russian

BABY PRODUCTS, English→Serbian					
		distribution			
system	score	m	f	x	mix
Algharb	*-83.0*	8	91	1	0
Gemini-2.5-Pro	*-83.0*	8	91	1	0
Wenyiil	*-74.0*	12	86	1	1
Shy	*-73.0*	13	86	1	0
ONLINE-B	*-72.0*	13	85	1	1
Yolu	*-65.0*	17	82	0	1
GemTrans	*-60.0*	19	79	1	1
hybrid	-57.0	20	77	1	2
Claude-4	-55.0	22	77	1	0
GPT-4.1	-52.0	23	75	1	1
IRB-MT	-51.0	22	73	1	4
Gemma-3-27B	-44.0	27	71	1	1
CUNI-SFT	-39.0	28	67	2	3
Gemma-3-12B	-39.0	26	65	1	8
EuroLLM-22B	-33.0	31	64	1	4
UvA-MT	-33.0	32	65	1	2
DeepSeek-V3	-23.0	37	60	1	2
AyaExpanse-32B	-20.0	36	56	1	7
Llama-3.1-8B	-13.0	39	52	2	7
Qwen3-235B	-10.0	44	54	1	1
TowerPlus-9B	-9.0	39	48	3	10
IR-MultiagentMT	8.0	52	44	1	3
CommandA	9.0	51	42	1	6
Qwen2.5-7B	11.0	44	33	4	19
AyaExpanse-8B	12.0	47	35	0	18
Llama-4-Maverick	16.0	56	40	2	2
SalamandraTA	19.0	52	33	2	13
CommandA-MT	25.0	60	35	1	4
CommandR7B	25.0	54	29	7	10
EuroLLM-9B	26.0	61	35	1	3
TowerPlus-72B	32.0	62	30	0	8
ONLINE-G	63.0	70	7	2	21
Mistral-7B	70.0	77	7	1	15
TranssionMT	82.0	83	1	1	15
TranssionTranslate	95.0	95	0	2	3

Table 16: BABY PRODUCTS, Serbian

BEAUTY, English→Russian					
		distribution			
system	score	m	f	x	mix
ONLINE-G	*-79.0*	4	83	2	11
ONLINE-W	-78.0	7	85	1	7
Yandex	*-75.0*	11	86	3	0
Algharb	*-74.0*	10	84	6	0
GemTrans	-74.0	11	85	2	2
Claude-4	-73.0	11	84	5	0
Gemini-2.5-Pro	*-73.0*	11	84	4	1
Yolu	*-73.0*	11	84	3	2
hybrid	-71.0	12	83	5	0
Wenyiil	*-71.0*	11	82	4	3
Shy	*-69.0*	13	82	5	0
IRB-MT	-67.0	14	81	3	2
Gemma-3-12B	-66.0	15	81	2	2
GPT-4.1	-64.0	16	80	4	0
DeepSeek-V3	-62.0	16	78	5	1
ONLINE-B	-62.0	16	78	1	5
DLUT_GTCOM	-60.0	19	79	2	0
Laniqo	-60.0	16	76	7	1
UvA-MT	-59.0	19	78	1	2
Gemma-3-27B	-58.0	19	77	2	2
TowerPlus-9B	-53.0	22	75	2	1
CommandA	-51.0	22	73	2	3
CommandA-MT	-50.0	24	74	1	1
Qwen3-235B	-49.0	24	73	2	1
TranssionTranslate	-49.0	16	65	1	18
EuroLLM-22B	-48.0	24	72	2	2
AyaExpanse-32B	-45.0	26	71	1	2
SalamandraTA	-45.0	26	71	1	2
TowerPlus-72B	-44.0	26	70	1	3
IR-MultiagentMT	-34.0	31	65	3	1
Llama-4-Maverick	-34.0	31	65	2	2
AyaExpanse-8B	-26.0	35	61	3	1
SRPOL	-19.0	38	57	3	2
Qwen2.5-7B	-18.0	32	50	8	10
Llama-3.1-8B	-13.0	38	51	2	9
CommandR7B	-10.0	42	52	1	5
EuroLLM-9B	-2.0	47	49	3	1
NLLB	70.0	81	11	8	0
Mistral-7B	80.0	87	7	2	4
TranssionMT	99.0	99	0	1	0

Table 17: BEAUTY AND PERSONAL CARE, Russian

BEAUTY, English→Serbian					
		distribution			
system	score	m	f	x	mix
Yolu	*-86.0*	6	92	0	2
ONLINE-B	*-84.0*	6	90	0	4
GemTrans	*-81.0*	9	90	1	0
Algharb	*-79.0*	10	89	0	1
Gemini-2.5-Pro	*-77.0*	11	88	0	1
GPT-4.1	-74.0	13	87	0	0
Shy	*-74.0*	12	86	0	2
Wenyiil	*-74.0*	11	85	0	4
hybrid	-72.0	14	86	0	0
IRB-MT	-69.0	15	84	1	0
Gemma-3-12B	-68.0	15	83	0	2
Claude-4	-66.0	16	82	0	2
UvA-MT	-60.0	19	79	0	2
DeepSeek-V3	-58.0	20	78	0	2
EuroLLM-22B	-53.0	21	74	1	4
Gemma-3-27B	-53.0	22	75	0	3
CUNI-SFT	-41.0	26	67	0	7
Qwen3-235B	-41.0	29	70	0	1
AyaExpanse-32B	-28.0	33	61	0	6
IR-MultiagentMT	-27.0	35	62	0	3
CommandA	-25.0	35	60	0	5
TowerPlus-9B	-21.0	33	54	3	10
EuroLLM-9B	-19.0	37	56	0	7
AyaExpanse-8B	-18.0	30	48	1	21
CommandA-MT	-7.0	45	52	0	3
Llama-4-Maverick	-4.0	46	50	0	4
Llama-3.1-8B	**0.0**	45	45	1	9
SalamandraTA	2.0	47	45	0	8
CommandR7B	3.0	41	38	6	15
Qwen2.5-7B	25.0	55	30	1	14
TowerPlus-72B	40.0	66	26	0	8
Mistral-7B	72.0	80	8	1	11
ONLINE-G	74.0	75	1	0	24
TranssionMT	85.0	86	1	0	13
TranssionTranslate	91.0	91	0	0	9

Table 18: BEAUTY AND PERSONAL CARE, Serbian

HEALTH, English→Russian					
		distribution			
system	score	m	f	x	mix
ONLINE-G	*-53.0*	11	64	1	24
Yandex	*-39.0*	29	68	2	1
Algharb	*-36.0*	30	66	1	3
Gemini-2.5-Pro	-31.0	34	65	1	0
Yolu	*-30.0*	34	64	1	1
Laniqo	-26.0	34	60	4	2
ONLINE-W	-25.0	23	48	1	28
Wenyiil	*-20.0*	39	59	2	0
Shy	*-15.0*	41	56	1	2
IRB-MT	-14.0	41	55	2	2
GemTrans	-9.0	44	53	2	1
DLUT_GTCOM	-4.0	46	50	1	3
Gemma-3-12B	-1.0	48	49	1	2
DeepSeek-V3	3.0	49	46	4	1
SalamandraTA	5.0	49	44	2	5
TowerPlus-9B	5.0	52	47	1	0
Claude-4	14.0	56	42	2	0
hybrid	14.0	54	40	4	2
UvA-MT	15.0	57	42	1	0
EuroLLM-22B	19.0	58	39	2	1
AyaExpanse-32B	23.0	60	37	1	2
TranssionTranslate	27.0	54	27	1	18
Qwen3-235B	28.0	63	35	1	1
GPT-4.1	29.0	64	35	1	0
CommandA	35.0	66	31	1	2
Qwen2.5-7B	44.0	66	22	7	5
Gemma-3-27B	49.0	73	24	1	2
Llama-3.1-8B	49.0	72	23	0	5
TowerPlus-72B	50.0	74	24	1	1
AyaExpanse-8B	51.0	73	22	1	4
SRPOL	51.0	75	24	1	0
ONLINE-B	52.0	74	22	1	3
IR-MultiagentMT	53.0	73	20	6	1
CommandR7B	61.0	79	18	1	2
Llama-4-Maverick	69.0	84	15	1	0
CommandA-MT	72.0	85	13	1	1
EuroLLM-9B	84.0	91	7	1	1
NLLB	87.0	91	4	3	2
Mistral-7B	91.0	94	3	0	3
TranssionMT	98.0	98	0	1	1

Table 19: HEALTH AND HOUSEHOLD, Russian

HEALTH, English→Serbian					
		distribution			
system	score	m	f	x	mix
ONLINE-B	*-43.0*	27	70	0	3
Gemini-2.5-Pro	*-36.0*	32	68	0	0
Algharb	*-32.0*	33	65	0	2
GemTrans	*-30.0*	34	64	0	2
Shy	*-23.0*	38	61	0	1
Yolu	*-22.0*	38	60	0	2
Wenyiil	*-15.0*	42	57	0	1
IRB-MT	-12.0	43	55	0	2
CUNI-SFT	-5.0	44	49	0	7
Gemma-3-12B	8.0	50	42	1	7
DeepSeek-V3	11.0	55	44	0	1
AyaExpanse-8B	13.0	48	35	0	17
Claude-4	13.0	56	43	0	1
hybrid	14.0	56	42	0	2
EuroLLM-22B	15.0	54	39	0	7
GPT-4.1	15.0	57	42	0	1
UvA-MT	19.0	59	40	0	1
Gemma-3-27B	20.0	57	37	0	6
AyaExpanse-32B	23.0	59	36	0	5
Llama-3.1-8B	31.0	61	30	0	9
Qwen3-235B	33.0	66	33	0	1
TowerPlus-9B	33.0	60	27	2	11
CommandR7B	34.0	58	24	6	12
CommandA	50.0	74	24	0	2
SalamandraTA	58.0	73	15	0	12
IR-MultiagentMT	59.0	79	20	0	1
EuroLLM-9B	60.0	75	15	0	10
Llama-4-Maverick	64.0	81	17	0	2
Qwen2.5-7B	70.0	80	10	1	9
TowerPlus-72B	75.0	85	10	0	5
CommandA-MT	84.0	92	8	0	0
Mistral-7B	85.0	89	4	0	7
TranssionMT	85.0	87	2	0	11
ONLINE-G	88.0	89	1	0	10
TranssionTranslate	99.0	99	0	0	1

Table 20: HEALTH AND HOUSEHOLD, Serbian

HOME AND KITCHEN, English→Russian					
		distribution			
system	score	m	f	x	mix
ONLINE-G	*-49.0*	15	64	2	19
Yolu	*-45.0*	27	72	1	0
Yandex	*-31.0*	34	65	1	0
Algharb	*-29.0*	34	63	1	2
Gemini-2.5-Pro	*-22.0*	38	60	1	1
GemTrans	-17.0	41	58	1	0
Wenyiil	*-16.0*	41	57	1	1
ONLINE-W	-11.0	29	40	2	29
Shy	*-10.0*	43	53	2	2
IRB-MT	-5.0	47	52	1	0
TowerPlus-9B	-5.0	47	52	1	0
Gemma-3-12B	-3.0	48	51	1	0
Claude-4	-2.0	48	50	1	1
Laniqo	-2.0	47	49	2	2
DLUT_GTCOM	**0.0**	48	48	2	2
hybrid	1.0	50	49	1	0
UvA-MT	4.0	51	47	2	0
GPT-4.1	7.0	53	46	1	0
SalamandraTA	13.0	55	42	1	2
DeepSeek-V3	19.0	59	40	1	0
AyaExpanse-32B	22.0	60	38	1	1
EuroLLM-22B	24.0	60	36	1	3
Gemma-3-27B	24.0	61	37	1	1
Qwen3-235B	29.0	63	34	1	2
TranssionTranslate	31.0	53	22	2	23
AyaExpanse-8B	32.0	65	33	1	1
TowerPlus-72B	34.0	66	32	1	1
ONLINE-B	40.0	66	26	1	7
IR-MultiagentMT	43.0	69	26	4	1
CommandA	44.0	71	27	1	1
Qwen2.5-7B	46.0	61	15	11	13
SRPOL	49.0	74	25	1	0
CommandR7B	53.0	74	21	1	4
Llama-3.1-8B	58.0	76	18	0	6
CommandA-MT	59.0	79	20	1	0
Llama-4-Maverick	62.0	80	18	1	1
EuroLLM-9B	70.0	84	14	1	1
NLLB	84.0	91	7	2	0
Mistral-7B	87.0	92	5	0	3
TranssionMT	99.0	99	0	0	1

Table 21: HOME AND KITCHEN, Russian

HOME AND KITCHEN, English→Serbian					
		distribution			
system	score	m	f	x	mix
Algharb	*-35.0*	32	67	0	1
Gemini-2.5-Pro	*-32.0*	34	66	0	0
ONLINE-B	*-26.0*	34	60	0	6
Yolu	*-26.0*	36	62	1	1
GemTrans	*-17.0*	40	57	0	3
Shy	*-12.0*	44	56	0	0
Wenyiil	*-12.0*	44	56	0	0
GPT-4.1	-6.0	47	53	0	0
CUNI-SFT	-5.0	44	49	0	7
Claude-4	**0.0**	50	50	**0**	**0**
hybrid	14.0	56	42	1	1
EuroLLM-22B	15.0	55	40	0	5
IRB-MT	18.0	58	40	0	2
Gemma-3-12B	24.0	53	29	0	18
TowerPlus-9B	27.0	56	29	5	10
AyaExpanse-32B	29.0	63	34	0	3
Gemma-3-27B	30.0	61	31	1	7
Qwen3-235B	32.0	66	34	0	0
UvA-MT	32.0	65	33	0	2
AyaExpanse-8B	35.0	62	27	0	11
DeepSeek-V3	36.0	68	32	0	0
Llama-3.1-8B	39.0	67	28	0	5
CommandR7B	41.0	60	19	4	17
IR-MultiagentMT	49.0	74	25	0	1
Qwen2.5-7B	51.0	70	19	0	11
CommandA	54.0	76	22	0	2
EuroLLM-9B	59.0	77	18	0	5
SalamandraTA	62.0	77	15	0	8
Llama-4-Maverick	69.0	84	15	0	1
CommandA-MT	75.0	86	11	0	3
TowerPlus-72B	77.0	87	10	0	3
ONLINE-G	82.0	82	0	0	18
Mistral-7B	88.0	91	3	0	6
TranssionMT	94.0	94	0	0	6
TranssionTranslate	99.0	99	0	0	1

Table 22: HOME AND KITCHEN, Serbian

MUSICAL INSTRUMENTS, English→Russian					
		distribution			
system	score	m	f	x	mix
SalamandraTA	69.0	83	14	1	2
Laniqo	70.0	83	13	4	0
TowerPlus-9B	70.0	85	15	0	0
Yandex	*72.0*	86	14	0	0
Algharb	*74.0*	87	13	0	0
Gemini-2.5-Pro	*74.0*	86	12	2	0
Wenyiil	*75.0*	87	12	1	0
ONLINE-G	*76.0*	82	6	1	11
Shy	*76.0*	87	11	2	0
DeepSeek-V3	77.0	86	9	4	1
Yolu	*78.0*	88	10	1	1
hybrid	80.0	88	8	4	0
Claude-4	81.0	90	9	1	0
IRB-MT	81.0	90	9	1	0
AyaExpanse-32B	82.0	91	9	0	0
AyaExpanse-8B	82.0	91	9	0	0
GemTrans	82.0	91	9	0	0
GPT-4.1	83.0	91	8	1	0
CommandA	85.0	92	7	1	0
Gemma-3-12B	85.0	92	7	1	0
DLUT_GTCOM	87.0	92	5	3	0
Qwen3-235B	87.0	93	6	1	0
SRPOL	87.0	93	6	1	0
Llama-4-Maverick	88.0	94	6	0	0
TowerPlus-72B	88.0	94	6	0	0
CommandR7B	89.0	94	5	1	0
NLLB	89.0	91	2	6	1
EuroLLM-22B	90.0	95	5	0	0
UvA-MT	90.0	95	5	0	0
IR-MultiagentMT	91.0	95	4	0	1
Qwen2.5-7B	91.0	93	2	2	3
CommandA-MT	92.0	96	4	0	0
EuroLLM-9B	93.0	96	3	1	0
Gemma-3-27B	93.0	96	3	1	0
ONLINE-W	93.0	93	0	1	6
Llama-3.1-8B	94.0	95	1	1	3
TranssionTranslate	94.0	94	0	1	5
ONLINE-B	95.0	96	1	1	2
Mistral-7B	97.0	98	1	1	0
TranssionMT	99.0	99	0	1	0

Table 23: MUSICAL INSTRUMENTS, Russian

MUSICAL INSTRUMENTS, English→Serbian					
		distribution			
system	score	m	f	x	mix
CUNI-SFT	59.0	77	18	0	5
AyaExpanse-8B	69.0	81	12	1	6
AyaExpanse-32B	70.0	84	14	0	2
Yolu	*70.0*	84	14	0	2
CommandR7B	72.0	79	7	10	4
Gemini-2.5-Pro	*72.0*	86	14	0	0
TowerPlus-9B	72.0	83	11	0	6
EuroLLM-22B	73.0	84	11	0	5
GemTrans	*74.0*	87	13	0	0
Algharb	*77.0*	88	11	0	1
Claude-4	78.0	89	11	0	0
Shy	*78.0*	89	11	0	0
ONLINE-B	*80.0*	89	9	0	2
Gemma-3-12B	82.0	89	7	0	4
Qwen3-235B	82.0	90	8	0	2
UvA-MT	82.0	91	9	0	0
hybrid	83.0	91	8	0	1
Qwen2.5-7B	84.0	91	7	0	2
SalamandraTA	85.0	91	6	0	3
CommandA	86.0	93	7	0	0
DeepSeek-V3	86.0	93	7	0	0
Gemma-3-27B	86.0	93	7	0	0
GPT-4.1	86.0	93	7	0	0
Wenyiil	*86.0*	93	7	0	0
EuroLLM-9B	89.0	93	4	0	3
IRB-MT	88.0	93	5	0	2
Llama-3.1-8B	88.0	93	5	0	2
IR-MultiagentMT	89.0	94	5	1	0
Llama-4-Maverick	89.0	94	5	0	1
TowerPlus-72B	92.0	96	4	0	0
CommandA-MT	94.0	97	3	0	0
Mistral-7B	94.0	97	3	0	0
TranssionMT	94.0	94	0	0	6
ONLINE-G	98.0	98	0	0	2
TranssionTranslate	100.0	100	0	0	0

Table 24: MUSICAL INSTRUMENTS, Serbian

PET SUPPLIES, English→Russian					
		distribution			
system	score	m	f	x	mix
Yandex	*-63.0*	17	80	3	0
Algharb	*-61.0*	19	80	1	0
Gemini-2.5-Pro	*-52.0*	24	76	0	0
Yolu	*-50.0*	24	74	2	0
Shy	*-41.0*	28	69	2	1
Wenyiil	*-39.0*	30	69	1	0
GemTrans	-37.0	31	68	1	0
TowerPlus-9B	-35.0	32	67	0	1
GPT-4.1	-29.0	35	64	1	0
ONLINE-G	*-29.0*	20	49	1	30
Claude-4	-22.0	38	60	2	0
hybrid	-18.0	38	56	2	4
IRB-MT	-18.0	40	58	2	0
DLUT_GTCOM	-15.0	40	55	1	4
DeepSeek-V3	-12.0	43	55	2	0
Laniqo	-9.0	44	53	1	2
SalamandraTA	-9.0	42	51	1	6
AyaExpanse-32B	-6.0	47	53	0	0
UvA-MT	-6.0	46	52	2	0
Gemma-3-12B	-5.0	46	51	2	1
Qwen3-235B	2.0	50	48	1	1
Gemma-3-27B	11.0	55	44	1	0
ONLINE-W	11.0	35	24	0	41
TowerPlus-72B	13.0	56	43	0	1
CommandA	15.0	57	42	1	0
EuroLLM-22B	16.0	56	40	1	3
Qwen2.5-7B	24.0	53	29	6	12
TranssionTranslate	29.0	51	22	0	27
Llama-4-Maverick	33.0	66	33	1	0
IR-MultiagentMT	35.0	65	30	3	2
AyaExpanse-8B	38.0	68	30	1	1
CommandR7B	38.0	66	28	3	3
ONLINE-B	38.0	68	30	0	2
Llama-3.1-8B	43.0	68	25	0	7
CommandA-MT	50.0	74	24	0	2
SRPOL	53.0	73	20	1	6
EuroLLM-9B	60.0	78	18	2	2
NLLB	76.0	86	10	1	3
Mistral-7B	93.0	93	0	0	7
TranssionMT	98.0	98	0	0	2

Table 25: PET SUPPLIES, Russian

PET SUPPLIES, English→Serbian					
		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	*-66.0*	17	83	0	0
Algharb	*-61.0*	19	80	0	1
ONLINE-B	*-47.0*	25	72	0	3
Wenyiil	*-47.0*	26	73	0	1
Shy	*-43.0*	28	71	0	1
GemTrans	*-39.0*	29	68	0	3
Yolu	*-39.0*	28	67	0	5
GPT-4.1	-19.0	40	59	0	1
hybrid	-19.0	39	58	1	2
Claude-4	-6.0	46	52	0	2
AyaExpanse-32B	8.0	50	42	0	8
Gemma-3-27B	8.0	52	44	0	4
IRB-MT	11.0	54	43	0	3
EuroLLM-22B	12.0	53	41	0	6
DeepSeek-V3	15.0	57	42	0	1
CUNI-SFT	16.0	54	38	0	8
TowerPlus-9B	17.0	50	33	4	13
UvA-MT	17.0	57	40	0	3
Gemma-3-12B	19.0	54	35	0	11
Llama-3.1-8B	21.0	58	37	0	5
Qwen3-235B	32.0	66	34	0	0
AyaExpanse-8B	33.0	60	27	2	11
CommandR7B	38.0	58	20	2	20
Qwen2.5-7B	40.0	62	22	1	15
IR-MultiagentMT	45.0	72	27	0	1
CommandA	48.0	71	23	0	6
SalamandraTA	48.0	71	23	0	6
Llama-4-Maverick	49.0	73	24	0	3
EuroLLM-9B	56.0	73	17	0	10
TowerPlus-72B	60.0	77	17	2	4
CommandA-MT	85.0	92	7	0	1
ONLINE-G	89.0	89	0	0	11
Mistral-7B	90.0	92	2	1	5
TranssionMT	93.0	94	1	0	5
TranssionTranslate	100.0	100	0	0	0

Table 26: PET SUPPLIES, Serbian

SPORTS AND OUTDOORS, English→Russian					
		distribution			
system	score	m	f	x	mix
SalamandraTA	5.0	49	44	1	6
Algharb	*8.0*	53	45	1	1
Yolu	*12.0*	55	43	1	1
Gemini-2.5-Pro	*17.0*	58	41	1	0
Wenyiil	*17.0*	57	40	1	2
Laniqo	20.0	59	39	2	0
Shy	*22.0*	59	37	3	1
Yandex	*22.0*	59	37	2	2
ONLINE-G	*24.0*	51	27	0	22
TowerPlus-9B	33.0	66	33	1	0
Claude-4	34.0	65	31	2	2
GemTrans	35.0	67	32	0	1
hybrid	35.0	65	30	4	1
GPT-4.1	37.0	68	31	1	0
EuroLLM-22B	41.0	70	29	0	1
IRB-MT	41.0	70	29	1	0
SRPOL	44.0	71	27	1	1
AyaExpanse-32B	45.0	72	27	0	1
CommandA	46.0	73	27	0	0
DeepSeek-V3	46.0	72	26	2	0
Gemma-3-12B	46.0	72	26	0	2
DLUT_GTCOM	47.0	71	24	3	2
UvA-MT	47.0	73	26	0	1
Qwen3-235B	49.0	74	25	0	1
IR-MultiagentMT	53.0	76	23	1	0
TowerPlus-72B	54.0	76	22	0	2
Gemma-3-27B	56.0	77	21	1	1
ONLINE-W	56.0	66	10	0	24
CommandR7B	58.0	76	18	3	3
AyaExpanse-8B	60.0	79	19	0	2
Llama-3.1-8B	60.0	79	19	0	2
Qwen2.5-7B	60.0	75	15	9	1
ONLINE-B	62.0	78	16	1	5
CommandA-MT	65.0	82	17	0	1
Llama-4-Maverick	65.0	82	17	1	0
EuroLLM-9B	66.0	83	17	0	0
TranssionTranslate	69.0	78	9	0	13
NLLB	80.0	88	8	0	4
Mistral-7B	90.0	94	4	0	2
TranssionMT	98.0	98	0	0	2

Table 27: SPORTS AND OUTDOORS, Russian

SPORTS AND OUTDOORS, English→Serbian					
		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	*4.0*	52	48	0	0
Algharb	*6.0*	53	47	0	0
Wenyiil	*10.0*	55	45	0	0
EuroLLM-22B	16.0	56	40	0	4
Yolu	*22.0*	60	38	0	2
CUNI-SFT	26.0	58	32	0	10
Shy	*27.0*	63	36	0	1
GemTrans	*28.0*	63	35	0	2
ONLINE-B	*31.0*	64	33	0	3
GPT-4.1	34.0	67	33	0	0
Claude-4	36.0	68	32	0	0
AyaExpanse-32B	37.0	65	28	1	6
IRB-MT	37.0	68	31	0	1
hybrid	38.0	69	31	0	0
TowerPlus-9B	40.0	65	25	1	9
DeepSeek-V3	41.0	70	29	0	1
Gemma-3-12B	41.0	69	28	0	3
AyaExpanse-8B	42.0	65	23	1	11
CommandR7B	45.0	63	18	4	15
UvA-MT	45.0	72	27	0	1
Gemma-3-27B	48.0	72	24	0	4
IR-MultiagentMT	48.0	73	25	2	0
Llama-3.1-8B	53.0	73	20	0	7
Qwen3-235B	57.0	78	21	0	1
EuroLLM-9B	58.0	74	16	1	9
CommandA	60.0	79	19	0	2
Qwen2.5-7B	64.0	75	11	3	11
SalamandraTA	68.0	83	15	0	2
Llama-4-Maverick	69.0	84	15	0	1
CommandA-MT	70.0	84	14	0	2
TowerPlus-72B	71.0	84	13	0	3
Mistral-7B	81.0	85	4	0	11
ONLINE-G	89.0	91	2	0	7
TranssionMT	92.0	94	2	0	4
TranssionTranslate	100.0	100	0	0	0

Table 28: SPORTS AND OUTDOORS, Serbian

TOOLS, English→Russian					
		distribution			
system	score	m	f	x	mix
Yolu	*29.0*	63	34	1	2
Gemini-2.5-Pro	*33.0*	66	33	1	0
Yandex	*36.0*	68	32	0	0
Algharb	*38.0*	68	30	2	0
ONLINE-G	*39.0*	58	19	1	22
Laniqo	41.0	67	26	5	2
TowerPlus-9B	41.0	70	29	0	1
SalamandraTA	47.0	71	24	3	2
Wenyiil	*48.0*	73	25	2	0
Shy	*50.0*	74	24	2	0
GPT-4.1	54.0	76	22	2	0
hybrid	56.0	75	19	5	1
GemTrans	58.0	79	21	0	0
DeepSeek-V3	59.0	79	20	1	0
IRB-MT	59.0	79	20	1	0
AyaExpanse-32B	61.0	78	17	4	1
Gemma-3-12B	61.0	80	19	1	0
Claude-4	62.0	80	18	2	0
Qwen3-235B	62.0	80	18	2	0
EuroLLM-22B	64.0	80	16	3	1
TowerPlus-72B	66.0	82	16	1	1
UvA-MT	66.0	82	16	2	0
ONLINE-W	67.0	76	9	1	14
AyaExpanse-8B	68.0	84	16	0	0
CommandA	69.0	83	14	3	0
IR-MultiagentMT	69.0	82	13	5	0
Gemma-3-27B	70.0	85	15	0	0
SRPOL	70.0	84	14	1	1
DLUT_GTCOM	71.0	84	13	2	1
Llama-3.1-8B	72.0	82	10	0	8
Llama-4-Maverick	77.0	87	10	3	0
ONLINE-B	77.0	87	10	1	2
TranssionTranslate	77.0	84	7	1	8
CommandR7B	79.0	87	8	2	3
Qwen2.5-7B	80.0	87	7	2	4
CommandA-MT	83.0	91	8	1	0
EuroLLM-9B	85.0	91	6	3	0
NLLB	91.0	94	3	2	1
Mistral-7B	95.0	96	1	0	3
TranssionMT	99.0	99	0	0	1

Table 29: TOOLS AND HOME IMPROVEMENT, Russian

TOOLS, English→Serbian					
		distribution			
system	score	m	f	x	mix
Gemini-2.5-Pro	*32.0*	66	34	0	0
ONLINE-B	*33.0*	66	33	0	1
Algharb	*34.0*	67	33	0	0
CUNI-SFT	38.0	65	27	1	7
GemTrans	*38.0*	69	31	0	0
Yolu	*41.0*	70	29	0	1
EuroLLM-22B	43.0	69	26	0	5
Wenyiil	*44.0*	72	28	0	0
Shy	*46.0*	73	27	0	0
AyaExpanse-32B	47.0	70	23	0	7
GPT-4.1	51.0	75	24	0	1
CommandR7B	52.0	70	18	3	9
IRB-MT	52.0	76	24	0	0
TowerPlus-9B	52.0	72	20	2	6
Gemma-3-12B	53.0	75	22	0	3
Gemma-3-27B	53.0	76	23	1	0
hybrid	54.0	77	23	0	0
Claude-4	56.0	78	22	0	0
AyaExpanse-8B	57.0	73	16	0	11
UvA-MT	62.0	81	19	0	0
DeepSeek-V3	66.0	83	17	0	0
CommandA	71.0	84	13	0	3
IR-MultiagentMT	72.0	85	13	0	2
Llama-3.1-8B	72.0	85	13	0	2
Qwen3-235B	73.0	86	13	0	1
Qwen2.5-7B	74.0	81	7	0	12
EuroLLM-9B	75.0	81	6	0	13
Llama-4-Maverick	77.0	88	11	0	1
TowerPlus-72B	79.0	88	9	0	3
CommandA-MT	80.0	90	10	0	0
SalamandraTA	80.0	88	8	0	4
Mistral-7B	90.0	93	3	0	4
ONLINE-G	91.0	91	0	1	8
TranssionMT	92.0	92	0	0	8
TranssionTranslate	100.0	100	0	0	0

Table 30: TOOLS AND HOME IMPROVEMENT, Serbian

VIDEO GAMES, English→Russian					
		distribution			
system	score	m	f	x	mix
Yandex	*60.0*	79	19	2	0
Yolu	*63.0*	80	17	1	2
SalamandraTA	65.0	81	16	1	2
Laniqo	69.0	83	14	2	1
Shy	*75.0*	85	10	5	0
Gemini-2.5-Pro	*77.0*	87	10	3	0
TowerPlus-9B	77.0	88	11	1	0
Gemma-3-12B	78.0	87	9	3	1
Algharb	*79.0*	89	10	1	0
ONLINE-G	*79.0*	85	6	1	8
hybrid	80.0	88	8	4	0
IRB-MT	80.0	89	9	2	0
TowerPlus-72B	81.0	90	9	1	0
UvA-MT	82.0	91	9	0	0
Wenyiil	*83.0*	90	7	3	0
Claude-4	84.0	91	7	2	0
Qwen2.5-7B	84.0	88	4	7	1
Qwen3-235B	84.0	92	8	0	0
IR-MultiagentMT	85.0	91	6	3	0
DeepSeek-V3	86.0	92	6	2	0
ONLINE-W	87.0	90	3	3	4
EuroLLM-22B	88.0	93	5	1	1
Gemma-3-27B	88.0	93	5	2	0
GPT-4.1	88.0	93	5	2	0
SRPOL	89.0	94	5	0	1
AyaExpanse-32B	90.0	95	5	0	0
AyaExpanse-8B	90.0	95	5	0	0
GemTrans	90.0	95	5	0	0
NLLB	90.0	92	2	4	2
CommandR7B	92.0	95	3	1	1
CommandA	93.0	96	3	1	0
TranssionTranslate	93.0	94	1	2	3
CommandA-MT	94.0	96	2	2	0
Llama-4-Maverick	94.0	97	3	0	0
DLUT_GTCOM	95.0	97	2	1	0
EuroLLM-9B	97.0	98	1	1	0
ONLINE-B	98.0	98	0	2	0
Mistral-7B	99.0	99	0	0	1
TranssionMT	99.0	99	0	1	0
Llama-3.1-8B	100.0	100	0	0	0

Table 31: VIDEO GAMES, Russian

VIDEO GAMES, English→Serbian					
		distribution			
system	score	m	f	x	mix
CUNI-SFT	52.0	74	22	1	3
AyaExpanse-8B	57.0	74	17	0	9
AyaExpanse-32B	67.0	80	13	1	6
CommandR7B	71.0	80	9	6	5
EuroLLM-22B	71.0	83	12	0	5
TowerPlus-9B	73.0	83	10	1	6
Yolu	*77.0*	87	10	0	3
Algharb	*78.0*	89	11	0	0
Shy	*78.0*	89	11	0	0
Wenyiil	*78.0*	89	11	0	0
Claude-4	80.0	90	10	0	0
Gemini-2.5-Pro	*80.0*	90	10	0	0
Gemma-3-12B	80.0	88	8	0	4
CommandA	81.0	89	8	0	3
GemTrans	*81.0*	90	9	0	1
IRB-MT	81.0	90	9	0	1
DeepSeek-V3	84.0	92	8	0	0
EuroLLM-9B	87.0	92	5	0	3
Qwen2.5-7B	87.0	88	1	0	11
UvA-MT	87.0	93	6	0	1
GPT-4.1	88.0	94	6	0	0
ONLINE-B	*88.0*	94	6	0	0
Gemma-3-27B	89.0	92	3	0	5
Llama-3.1-8B	90.0	94	4	0	2
SalamandraTA	90.0	91	1	1	7
TowerPlus-72B	91.0	95	4	1	0
hybrid	92.0	96	4	0	0
Qwen3-235B	92.0	96	4	0	0
IR-MultiagentMT	94.0	96	2	0	2
TranssionMT	94.0	95	1	1	3
ONLINE-G	96.0	96	0	0	4
Mistral-7B	97.0	98	1	0	1
CommandA-MT	98.0	99	1	0	0
Llama-4-Maverick	98.0	99	1	0	0
TranssionTranslate	100.0	100	0	0	0

Table 32: VIDEO GAMES, Serbian

Evaluation of LLM for English to Hindi Legal Domain Machine Translation Systems

Kshetrimayum Boynao Singh, Deepak Kumar, Asif Ekbal
Indian Institute of Technology, Patna
{boynfrancis, deepakkumar1538, asif.ekbal}@gmail.com

Abstract

The study critically examines various Machine Translation systems, particularly focusing on Large Language Models, using the WMT25 Legal Domain Test Suite for translating English into Hindi. It utilizes a dataset of 5,000 sentences designed to capture the complexity of legal texts based on word frequency ranges from 5 to 54. Each frequency range contains 100 sentences, collectively forming a corpus that spans from simple legal terms to intricate legal provisions. Six metrics were used to evaluate the performance of the system: BLEU, METEOR, TER, CHRF++, BERTScore and COMET. The findings reveal diverse capabilities and limitations of LLM architectures in handling complex legal texts. Notably, Gemini-2.5-Pro, Claude-4, and ONLINE-B topped the performance charts in terms of human evaluation, showcasing the potential of LLMs for nuanced translation. Despite these advances, the study identified areas for further research, especially in improving robustness, reliability, and explainability for use in critical legal contexts. The study also supports the WMT25 subtask focused on evaluating the weaknesses of large language models (LLMs). The dataset and related resources are publicly available at https://github.com/helloboyn/WMT25-TS

1 Introduction

Machine Translation (MT) has evolved from basic rule-based and statistical approaches to advanced neural network models, with recent advancements driven by Large Language Models (LLMs) that utilize extensive pretraining datasets and transformer architectures. (Vaswani et al., 2017) The legal domain poses significant challenges for MT, requiring precise handling of context-dependent terminology (Appicharla et al., 2025), complex sentence structures, and the accurate conveyance of cultural and jurisdictional nuances due to varying legal systems. This complexity surpasses that found in general language translation, making high levels of lexical accuracy, logical coherence, and syntactic fidelity essential for effective legal translation.

The text highlights the crucial importance of accuracy and fidelity in legal document translation due to its high-stakes nature. It emphasizes that even minor mistranslations can lead to serious legal and financial consequences such as contractual disputes and judicial errors. Therefore, precise legal translation is essential to support international legal cooperation, manage cross-border litigation, provide equitable access to justice for non-native speakers, and make legal information accessible to a wider audience.(WMT)[1] The series has consistently served as a pivotal platform, instrumental in benchmarking progress and driving innovation within the machine translation research community across a diverse array of language pairs and domains. WMT25 (Kocmi et al., 2024) continues this vital tradition, offering meticulously designed specialized test suites that push the boundaries of current MT technologies and identify areas for future breakthroughs. This paper specifically focuses on the WMT25 Legal Domain Test Suite for English to Hindi [2], embarking on an in-depth investigation into how various LLM-based MT systems perform when compared to more traditional and established hybrid approaches. Our overarching objective is to provide a comprehensive and nuanced analysis of their efficacy in this demanding domain, meticulously identifying the top performing contenders and critically discussing the broader implications of our findings for the future trajectory and practical application of legal machine translation systems, including considerations for deployment and ethical use.

[1] https://www2.statmt.org/wmt25/testsuite-subtask.html
[2] https://github.com/wmt25testsuite/wmt25

Proceedings of the Tenth Conference on Machine Translation, pages 823–833
November 8-9, 2025

2 Related Work

The WMT shared tasks have consistently been a primary driving force behind significant advancements in Machine Translation research, fostering innovation and providing a standardized, competitive benchmark for evaluating system performance (Gain et al., 2022). Previous WMT editions, notably those from WMT24 (as evidenced by a series of influential papers such as (Freitag et al., 2024) to (Ármannsson et al., 2024)), have unequivocally showcased the increasing dominance and sophistication of neural MT (NMT) models (Appicharla et al., 2021). These works have meticulously detailed a wide array of architectural innovations, including the widespread adoption of transformer networks, advanced training methodologies such as back-translation and knowledge distillation, and impressive performance gains across diverse language pairs and specialized domains. Key themes emerging from this extensive body of research include the paramount importance of large-scale pre-training on vast textual corpora to learn robust linguistic representations, the efficacy of fine-tuning models on domain-specific data (Bhattacharjee et al., 2024; Moslem et al., 2022) to enhance specialized vocabulary and stylistic nuances (e.g., legal jargon, formal tone), and the continuous refinement of robust evaluation metrics to more accurately reflect human judgment of translation quality. Techniques like data augmentation (e.g., synthetic data generation), transfer learning (Singh et al., 2023a) from high-resource to low-resource languages, and the development of more efficient attention mechanisms have been central to these advancements, enabling NMT models (**?**) to capture intricate linguistic patterns and contextual dependencies with greater precision than their predecessors.

The rapid advancement of Large Language Models (LLMs) like GPT, Gemini, and Claude has significantly transformed Machine Translation (MT) research by challenging existing paradigms. These models, originally crafted for general language tasks, have shown impressive zero-shot and few-shot translation skills due to their training on vast, diverse datasets. They excel in capturing complex semantics and context, making them promising for specialized fields such as legal translation, where precision and adherence to terminology are critical. Research is actively exploring their (Gain et al., 2021) adaptation for specific MT tasks, and it often outperforms traditional Neural Machine Translation (NMT) (Singh et al., 2023b, 2024) models in challenging situations, such as low-resource languages and complex linguistic features (Manakhimova et al., 2024). Nonetheless, adapting LLMs to specific domains poses challenges, such as the risk of losing general linguistic knowledge, generating plausible but incorrect legal outputs, and maintaining strict legal fidelity without creative rephrasing.

3 Methodology

3.1 Dataset

The research uses a specialized Legal Domain Test Suite for WMT25 to evaluate translation systems from English to Hindi. This dataset consists of 5000 sentences derived from authentic legal documents on Table 1, reflecting the complexity and diversity of legal texts. It includes sentences varying in length from about 5 words to 54 words.

Word-Count	Sentences	Eng-Token	Hin-Token
5–15	1,600	20000	23046
16–35	1,700	49300	40798
36–54	1,700	78199	69627
Total	5,000	147499	133471

Table 1: Corpus statistics for the English and Hindi legal dataset by word count range.

The variation enables testing of systems' adaptability and robustness across different linguistic complexities, from precise legal terms to complex legal judgments. The dataset tests systems on their ability to handle the unique vocabulary, tone, and structure of legal language, ensuring accurate translation that maintains legal intent and avoids ambiguity.

3.2 Automatic Evaluation Metrics

To deliver a comprehensive and multifaceted (Chen et al., 2023) assessment of the translation quality generated by the various systems, six well-established and complementary automatic evaluation metrics were rigorously utilized. The choice of these metrics was deliberate, with the aim of capturing diverse aspects of translation quality: lexical overlap, semantic equivalence, and character-level accuracy, all of which are essential for the rigorous legal domain.

- **BLEU (Bilingual Evaluation Understudy):** The text examines the BLEU metric used

in machine translation evaluation, noting its strengths in precision, simplicity, and efficiency, which contribute to its widespread adoption. However, it also identifies BLEU's (Papineni et al., 2002) limitations in assessing translation fluency, grammatical correctness, and semantic adequacy, as it emphasizes lexical similarity over meaning. This focus can result in high scores for outputs closely matching references while neglecting valid paraphrases or alternative translations, which is particularly problematic in fields like legal translation, where multiple correct phrasings may exist.

- **METEOR (Metric for Evaluation of Translation With Explicit Ordering):**METEOR improves upon BLEU by using linguistic features such as word stemming, synonymy matching, and chunk-based alignment to better assess translation quality. By focusing on fluency and semantic adequacy, METEOR (Banerjee and Lavie, 2005) aligns more closely with human evaluations. It handles lexical and syntactic variations while penalizing reordering errors, making it particularly effective for domains requiring high semantic precision, such as legal texts.

- **TER (Term Error Rate):** TER, or Translation Edit Rate (Snover et al., 2006), is a metric used to evaluate the quality of machine translation (MT) by measuring the number of edits required to transform an MT output into a perfect, human-quality reference translation. These edits typically include insertions, deletions, substitutions, and shifts of words or phrases. A lower TER score indicates that fewer edits were necessary, meaning the machine translation is closer to the human reference and, thus, of higher quality. Conversely, a higher TER score signifies that many edits were needed, indicating poorer quality of machine translation that deviates significantly from the human standard.

- **CHRF++ (Character n-gram F-score):** The text discusses the CHRF++ (Popović, 2017) metric, which evaluates translation quality by computing the F-score of character n-grams between candidate and reference translations. It is highly regarded for its strong correlation with human judgments and its ability to handle morphological variations and out-of-vocabulary words effectively. CHRF++ is particularly suited for languages with rich morphological systems, such as Hindi, as it captures subtle character-level differences crucial for accurate translations. This makes it especially valuable in legal translation, where precise fidelity and a lack of ambiguity are critical.

- **BERTScore (Bidirectional Encoder Representations from Transformers Score):** BERTScore is a metric that assesses the quality of AI-generated text by measuring the semantic similarity between the generated content and reference texts (Zhang et al., 2020). Unlike traditional metrics that rely on exact word overlap, BERTScore uses the BERT language model to create embeddings of words and sentences, capturing their contextual meaning. A high BERTScore suggests that the generated text successfully conveys similar information and meaning to the reference, indicating good quality, while a low score points to significant differences in meaning, reflecting poor generation quality.

- **COMET (Crosslingual Optimized Metric for Evaluation of Translation):** COMET is an AI-based metric designed to evaluate the quality of machine translations by assessing alignment with high-quality human translations and considering the original source sentence for context. It uses a neural network model trained to align with human judgments, making it more robust and reliable than traditional rule-based metrics. High COMET (Rei et al., 2020) scores indicate superior translations, while low scores suggest poor translation quality.

3.3 Systems Evaluated

The WMT25 Legal Domain Test Suite served as a platform to evaluate a wide range of Machine Translation (MT) systems, reflecting the latest advancements in the field. It featured both proprietary and open-source Large Language Models (LLMs), such as Gemini-2.5-Pro, Claude-4, GPT-4.1, Llama, Mistral, Gemma, and Qwen, showcasing diverse architectures and scales. The evaluation (Manakhimova et al., 2023) also included traditional neural machine translation systems that

have been refined through years of domain adaptation and specialized training, as well as innovative hybrid approaches that incorporate rule-based systems or statistical models with neural components. These evaluations highlight the progress and state-of-the-art techniques in neural language processing, particularly in handling translations in the legal domain.

The paper conducts a comparative analysis of various translation systems—commercial, academic, and open-source large language models specifically within the legal domain. It assesses different model sizes and architectures, exploring the impact of scale, design, and training on translation quality and robustness. The study identifies the strengths and weaknesses of these systems, providing insights for future improvements and applications of machine translation in specialized areas.

4 Results and Observation

The performance of the systems evaluated on the WMT25 Legal Domain Test Suite (English to Hindi) is meticulously summarized below, directly derived from the provided evaluation results:

4.1 Overall Performance and LLM Dominance

The study highlights that LLM-based systems, especially Gemini-2.5-Pro, excel in machine translation within the legal domain, outperforming others across various metrics such as BLEU, METEOR, TER, CHRF++, BERTScore, and COMET. This is due to its extensive pre-training and specialized fine-tuning on legal documents, enhancing its handling of legal terminology and nuances. Other LLMs, such as Claude-4 and Llama-4-Maverick, also demonstrate strong performance, signaling a shift towards general-purpose models that outperform traditional systems in legal translation tasks. This shift offers legal professionals more efficient translation tools but also raises concerns about transparency and potential errors in precision-critical contexts.

4.2 Comparison with Non-LLM Systems

The text highlights that while Large Language Models (LLMs) are dominant in translation tasks, non-LLM or hybrid systems like ONLINE-B and TranssionTranslate also show competitive performance. Traditional Neural Machine Translation (NMT) systems, particularly those optimized for specific domains and language pairs, can achieve state-of-the-art results, offering computational efficiency and control over translation behavior. Hybrid approaches that integrate multiple translation paradigms enhance robustness and accuracy by combining the strengths of various methods, such as rule-based systems and statistical models. These alternatives are particularly viable in resource-limited settings demanding precision, such as legal translation, where accuracy and consistency are crucial.

4.3 Metric-Specific Observations

The evaluation of machine translation metrics highlights the varied strengths and weaknesses of different systems. The BLEU score primarily captures n-gram overlap, but its ability to assess semantic meaning and fluency is limited. Metrics such as BLEU, METEOR, TER, CHRF++, BERTScore, and COMET show different levels of effectiveness, with top models like Gemini-2.5-Pro, ONLINE-B, TranssionTranslate, ONLINE-G, and Claude-4 performing well overall (see Table 2). Gemini-2.5-Pro excels in precision and quality, while METEOR, focusing on semantic and structural accuracy, showcases ONLINE-G's strength. CHRF++ correlates well with translation quality through its character-level focus. A significant performance gap exists between leading and lower-tier models, with weaker systems performing poorly across metrics, indicating insufficient specialization in translation tasks. These metrics emphasize the strengths and constraints of each system in accurately translating legal texts.

4.4 Analysis by Sentence Length

The WMT25 Legal Domain Test Suite evaluates system performance over sentence lengths ranging from 5 to 54 words to assess robustness and adaptability to linguistic complexities. Although the dataset offers an aggregate performance overview, it lacks detailed scores segmented by sentence length. Such a breakdown is important for understanding how language models manage various contextual complexities and for identifying strengths or weaknesses related to sentence length. A more thorough evaluation would categorize sentences and assess performance within each segment.

Small Sentences (5–15 words) The analysis of short legal sentences shows that multiple machine translation (MT) systems excel in this area due to their minimal syntactic complexity. Systems such

Rank	**LLM System**	**BLEU**	**METEOR**	**TER**	**CHRF++**	**BERTScore**	**COMET**
1	Gemini-2.5-Pro	33.35	53.91	55.66	60.95	88.49	72.27
2	ONLINE-B	31.77	52.37	55.69	57.81	87.44	70.96
3	TranssionTranslate	31.65	52.42	55.71	57.83	87.55	71.01
4	ONLINE-G	31.22	57.30	52.07	55.20	86.56	67.37
5	Claude-4	31.09	52.75	57.87	58.46	87.71	70.99
6	Llama-4-Maverick	28.46	54.44	57.15	54.70	86.65	69.86
7	NLLB	27.87	51.55	57.45	53.38	86.11	68.16
8	hybrid	26.97	50.19	62.42	55.47	86.67	71.20
9	DeepSeek-V3	26.65	49.11	62.24	53.94	86.33	69.92
10	GPT-4.1	26.04	48.51	63.18	53.58	86.22	70.23
11	TowerPlus-9B	25.77	48.02	63.58	52.08	85.85	68.79
12	HYT	25.58	48.70	63.58	54	85.93	71.02
13	TMTHY	25.58	48.70	63.58	54	85.93	71.02
14	Shy	25.58	48.70	63.58	54	85.93	71.02
15	CommandA	24.24	47.85	65.11	51.70	85.64	68.85
16	Gemma-3-27B	23.80	46.22	65.91	51.49	85.35	68.96
17	TowerPlus-72B	23.53	46.57	65.42	50.24	85.32	67.76
18	Mistral-Medium	23.32	46.03	66.56	51.09	85.17	68.76
19	Qwen3-235B	22.91	45.75	66.96	50.33	85.14	68.20
20	EuroLLM-22B	22.18	44.72	67.39	48.94	84.76	67.40
21	EuroLLM-9B	21.52	44.65	68.68	48.09	84.35	66.23
22	Gemma-3-12B	21.51	43.81	68.04	49.38	84.47	68.03
23	IR-MultiagentMT	21.42	43.78	67.26	48.26	84.52	68.08
24	CommandA-MT	21.05	44.60	68.78	49.34	84.71	69.94
25	AyaExpanse-32B	20.50	43.75	69.00	47.50	84.28	66.91
26	UvA-MT	19.79	43.46	70.83	47.59	84.29	68
27	IRB-MT	17.26	39.92	84.85	44.15	83.63	67.21
28	AyaExpanse-8B	16.70	39.36	73.32	43.74	83.07	65.30
29	Llama-3.1-8B	15.21	38.54	74.68	42.20	82.03	63.19
30	GemTrans	15.16	38.68	80.62	43.06	82.05	67.76
31	CommandR7B	12.42	34.56	84.17	37.82	81.11	61.44
32	Qwen2.5-7B	8.75	27.88	87.18	33.22	78.52	53.05
33	Mistral-7B	3.03	20.65	177.39	23.19	71.04	41.79
34	Wenyiil	2.68	5.96	107.66	2.20	69.73	41.57
35	Yolu	2.68	5.96	107.66	2.20	69.73	41.57
36	Algharb	2.68	5.96	107.66	2.20	69.73	41.57
37	MMMT	2.68	5.96	107.66	2.20	69.73	41.57

Table 2: The table presents a performance comparison of various machine translation systems, including large language models (LLMs) and traditional neural machine translation (NMT) systems. We evaluate the systems using BLEU (Figure 1), METEOR (Figure 2), TER (Figure 3), CHRF++ (Figure 4), BERTScore (Figure 5), and COMET (Figure 6). The systems are ranked by their BLEU scores, with Gemini-2.5-Pro achieving the highest score, followed by ONLINE-B and TranssionTranslate. The results highlight the varying levels of translation quality across different models.

Rank	LLM System	Human Score%
1	Gemini-2.5-Pro	84.67
2	Claude-4	82.00
3	ONLINE-B	81.67
4	TowerPlus-9B	81.33
5	Llama-4-Maverick	81.00
6	GPT-4.1	80.67
7	TranssionTranslate	80.33
8	Qwen3-235B	80.00
9	Mistral-Medium	79.33
10	EuroLLM-22B	79.33
11	NLLB	78.67
12	HYT	78.67
13	ONLINE-G	78.33
14	DeepSeek-V3	78.33
15	TMTHY	78.33
16	CommandA	78.33
17	hybrid	78.00
18	Gemma-3-27B	78.00
19	Shy	77.67
20	TowerPlus-72B	76.74

Table 3: Human evaluation results for the top 20 BLEU-ranked systems on the English→Hindi legal domain dataset. Scores are averaged over two expert annotators.

as ONLINE-G, Llama-4-Maverick, and Claude-4 are identified as top performers, providing accurate and fluent translations. However, these models may face challenges with highly specialized legal jargon or rare terms not extensively covered in their training data, which could impact the precision required for translating legal documents.

Medium Sentences (16–35 words) The text discusses the challenges faced by translation models when dealing with medium-length sentences, which often contain complex structures, such as multiple clauses and conditional statements. These sentences require maintaining logical coherence and resolving anaphora for accurate translation. The models Gemini-2.5-Pro, TranssionTranslate, and ONLINE-B were identified as the most effective in managing these intricacies. Despite the models' suitability for this task, largely due to their transformer-based architectures, their performance showed a slight decline with shorter sentences, indicating that increased complexity still poses a risk of increased errors.

Large Sentences (36–54 words) The primary challenge for the 41 machine translation (MT) systems was translating long sentences characterized by legal jargon and numerous clauses. These sentences posed difficulties in maintaining contextual integrity, causing a significant drop in performance metrics such as BLEU and COMET. EuroLLM-22B, CommandA, and NLLB systems performed slightly better in mitigating this drop. The consistent performance decline underscores the increased risk of "hallucination," context loss, and ambiguity with longer sentences, marking it as a key area for future research and development in MT technology.

Overall System Performance

The evaluation ranks systems in Table 2 based on their robustness across different sentence lengths. Gemini-2.5-Pro is identified as the leading system, showing consistently high performance and the ability to manage various sentence complexities. It is followed by hybrid, with Shy, HYT, and TMTHY tied for third place. The assessment highlights that a system's overall performance is best measured by its consistent quality across all sentence lengths rather than excelling in just one category.

4.5 Human Analysis

In this section, we present the human evaluation conducted on the top 20 systems selected based on their BLEU scores in Table 3 . After identifying these top-performing systems, we carried out a detailed human evaluation specifically within the legal domain. Two linguistic experts proficient in both English and Hindi evaluated the translations of each system. They rated the outputs on a scale from 1 to 100, focusing on both adequacy (how accurately the translation conveyed the source meaning) and fluency (how natural and readable the translation was in Hindi).

We then averaged the scores from both evaluators to produce a final human evaluation score for each system. This human evaluation provides a nuanced measure of translation quality that complements the BLEU-based rankings, helping us identify systems that perform well in real-world, domain-specific scenarios.

5 Conclusion

The WMT25 Legal Domain Test Suite for English to Hindi Machine Translation highlights significant progress made by Large Language Models (LLMs) in specialized domain translation. Notably, Gemini-2.5-Pro excels, outperforming others across multiple evaluation metrics and emphasizing the potential of advanced LLM architectures with domain-specific pre-training or fine-tuning. These

models effectively handle complex legal language and structures, showcasing their sophisticated linguistic capabilities. While LLMs show dominance, traditional and hybrid MT systems also demonstrate competitiveness, indicating their continued relevance. The study underscores the importance of model scale, architecture, and domain adaptation for success in legal MT. It suggests that LLMs will play an increasingly central role in legal translation, advancing accuracy and efficiency. However, ongoing innovations across MT paradigms are needed to balance performance with reliability and ethical considerations, given the high stakes of errors in the legal domain.

Acknowledgement

The authors express their sincere gratitude to the COIL-D Project under Bhashini, funded by MeitY, for their support and resources, which were instrumental in the successful completion of this research.

References

Ramakrishna Appicharla, Asif Ekbal, and Pushpak Bhattacharyya. 2021. EduMT: Developing machine translation system for educational content in Indian languages. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 35–43, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2025. Beyond the sentence: A survey on context-aware machine translation with large language models. *arXiv preprint arXiv:2506.07583*.

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinþór Steingrímsson. 2024. Killing two flies with one stone: An attempt to break LLMs using English-Icelandic idioms and proper names. In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458, Miami, Florida, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Soham Bhattacharjee, Baban Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in English-Hindi translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 341–354, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted challenge set for evaluating machine translation performance. In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223, Singapore. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Baban Gain, Ramakrishna Appicharla, Soumya Chennabasavaraj, Nikesh Garera, Asif Ekbal, and Muthusamy Chelliah. 2022. Low resource chat translation: A benchmark for Hindi–English language pair. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 83–96, Orlando, USA. Association for Machine Translation in the Americas.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for Hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44, Online (Virtual Mode). INCOMA Ltd.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popovic, Mariya Shmatova, and 2 others. 2024. Preliminary wmt24 ranking of general mt systems and llms. *Preprint*, arXiv:2407.19884.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of*

the Ninth Conference on Machine Translation, pages 355–371, Miami, Florida, USA. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-specific text generation for machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLPAI).

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Evaluation Metrics and Dataset Segmentation

The appendix details benchmarking results for English to Hindi legal domain machine translation systems, evaluated using several metrics, including BLEU, METEOR, TER, chrF++, BERTScore and COMET. The performance of each system is analyzed across three sentence length categories—small, medium, and large—as well as an overall aggregate. Consistently high-performing systems are identified, along with those that rank lower in performance. The results show consistent top-performing systems, such as **Gemini-2.5-Pro**, **Claude-4**, and **TranssionTranslate**, while systems like **MMMT**, **Wenyill**, and **Yolu** consistently rank among the lowest. This segmentation provides deeper insights into system robustness across varying sentence complexities. Furthermore, it highlights the sensitivity of different models to sentence length, revealing cases in which certain systems degrade significantly with longer inputs. These findings underscore the importance of evaluating MT systems with controlled test suites to ensure reliability in specialized domains, such as legal translation.

A.1 Dataset Segmentation

The test data is divided into four buckets based on sentence length:

- **Small (5–15 words):** Marked in green.
- **Medium (16–35 words):** Marked in yellow.
- **Large (36–54 words):** Marked in blue.
- **Overall:** Aggregate results across all lengths, marked in red.

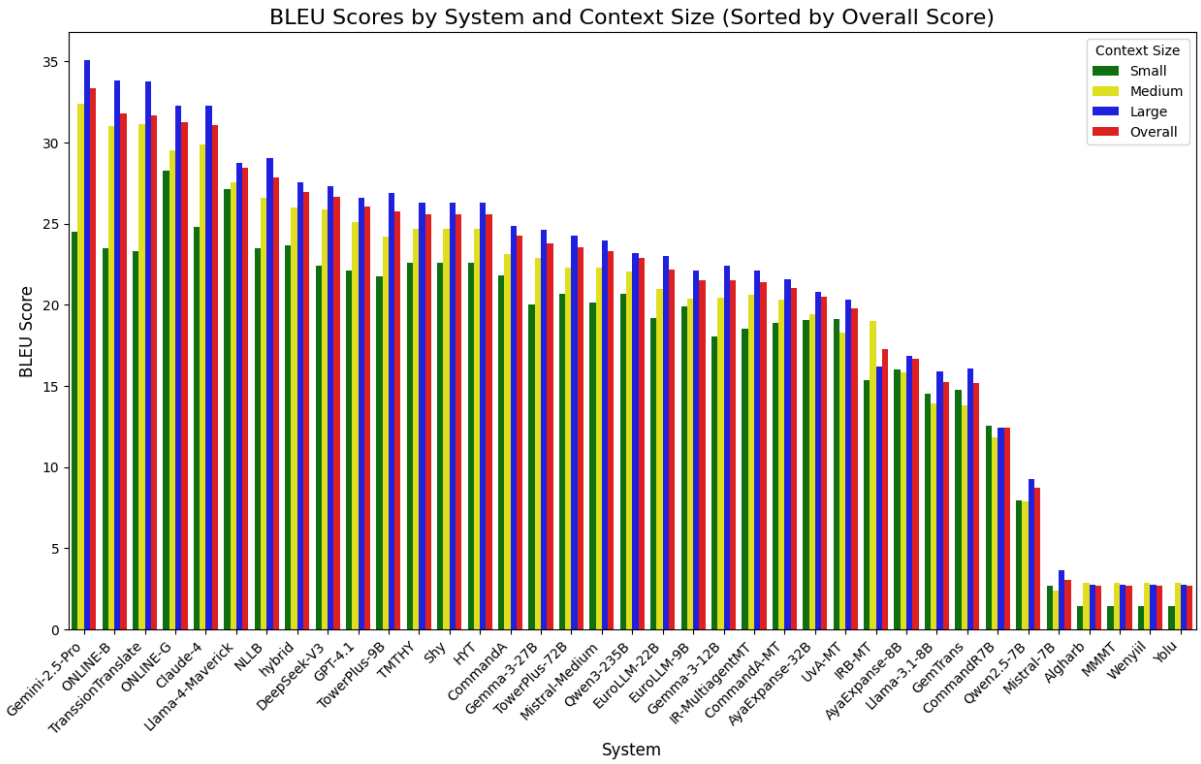


Figure 1: The bar chart displays the **BLEU scores** for various translation systems, broken down by **small, medium, and large context sizes**, as well as an **overall** score. The systems are ranked by their overall score, with **Gemini-2.5-Pro** and **Claude-4** having the highest overall BLEU score and **Wenyiil**, and **Yolu** having the lowest.

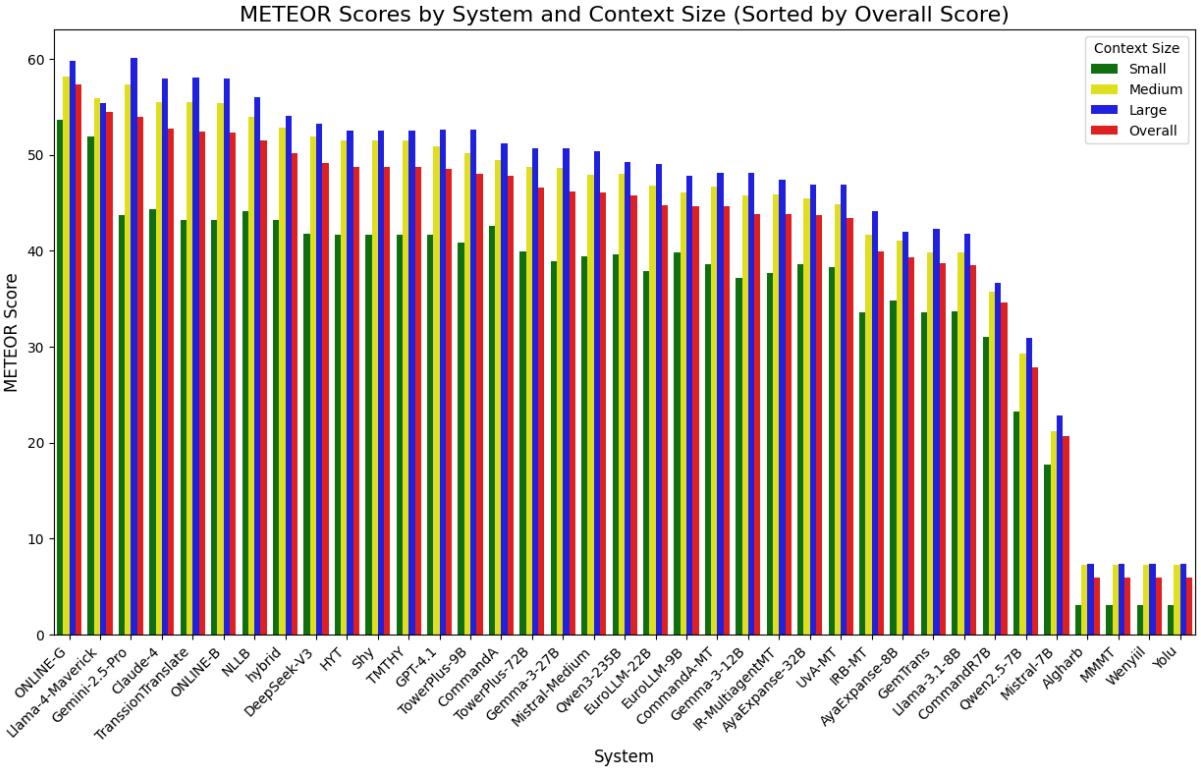


Figure 2: The bar chart shows **METEOR scores** for various translation systems, sorted by their **overall** performance. **ONLINE-G**, **Llama-4-Maverick** and **Gemini-2.5-Pro** have the highest scores, while **MMMT, Wenyiil, and Yolu** have the lowest.

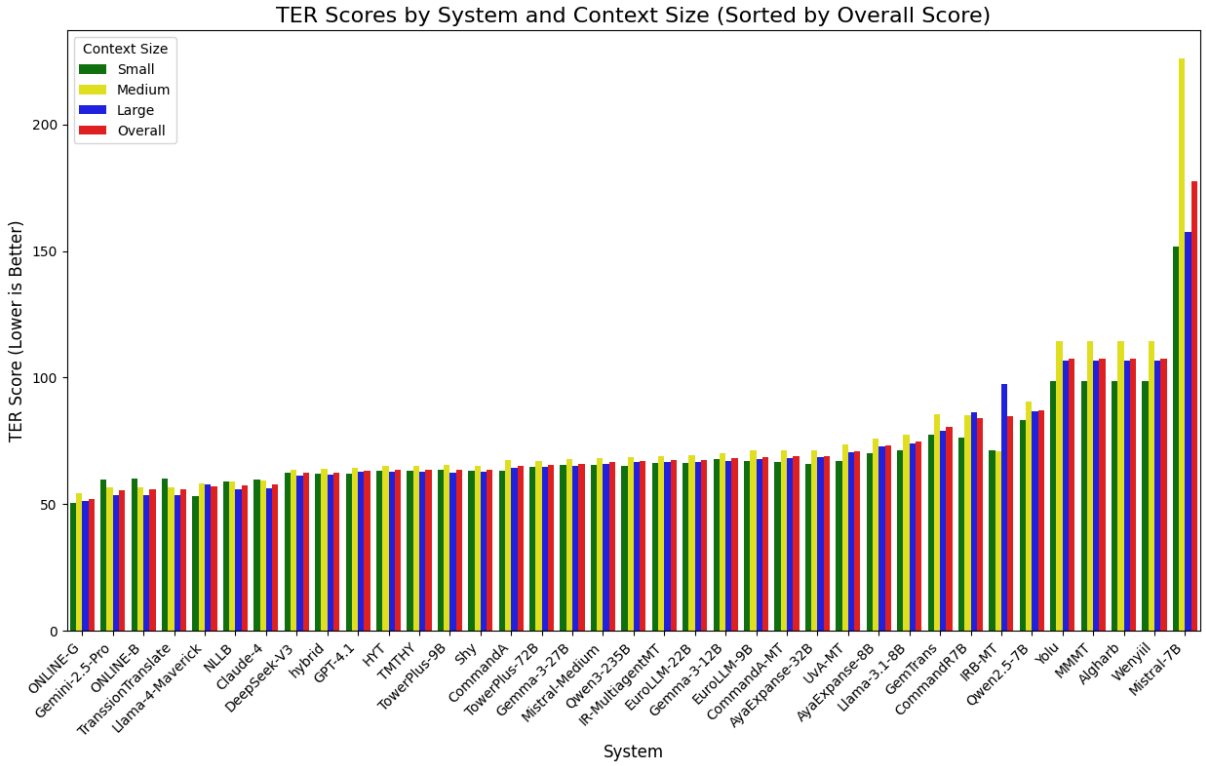


Figure 3: **TER scores** for various translation systems, sorted by their overall performance. **Lower scores are better**. **ONLINE-G, Gemini-2.5-Pro, and ONLINE-B** have the best performance, while **Mistral-7B, Wenyiil, and MMMT** have the worst.

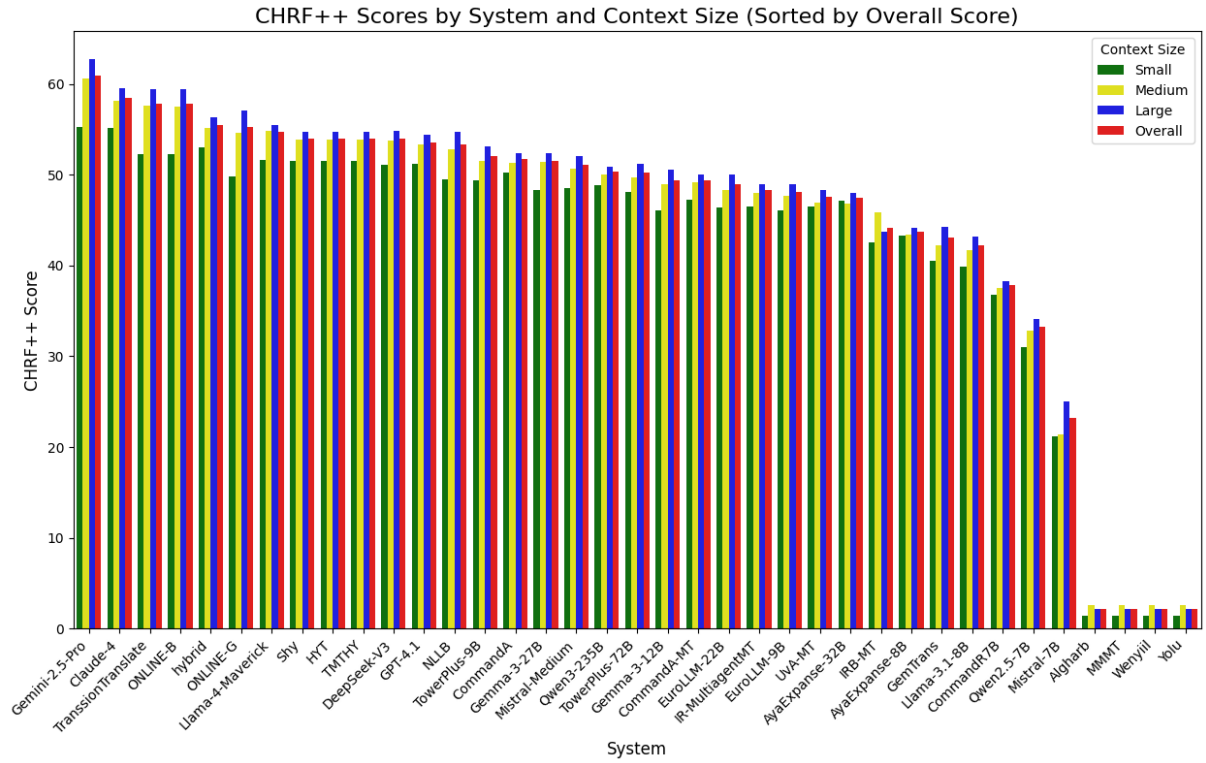


Figure 4: **CHRF++ scores** for various translation systems, sorted by their overall performance. **Higher scores are better**. **Gemini-2.5-Pro, Claude-4, and TranssionTranslate** have the best performance, while **MMMT, Wenyiil, and Yolu** have the worst.

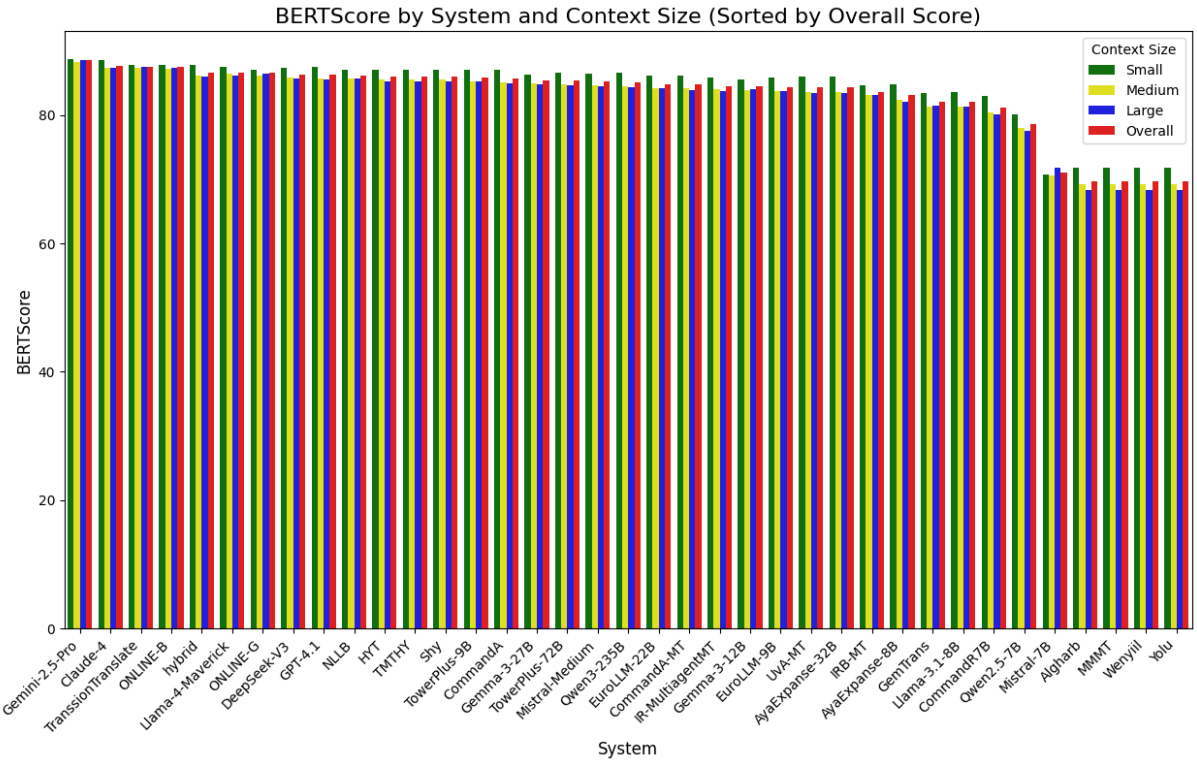


Figure 5: **BERTScore** for various translation systems, sorted by their overall performance. **Higher scores are better**. **Gemini-2.5-Pro, Claude-4, and TranssionTranslate** have the best performance, while **MMMT, Wenyiil, and Yolu** have the lowest.

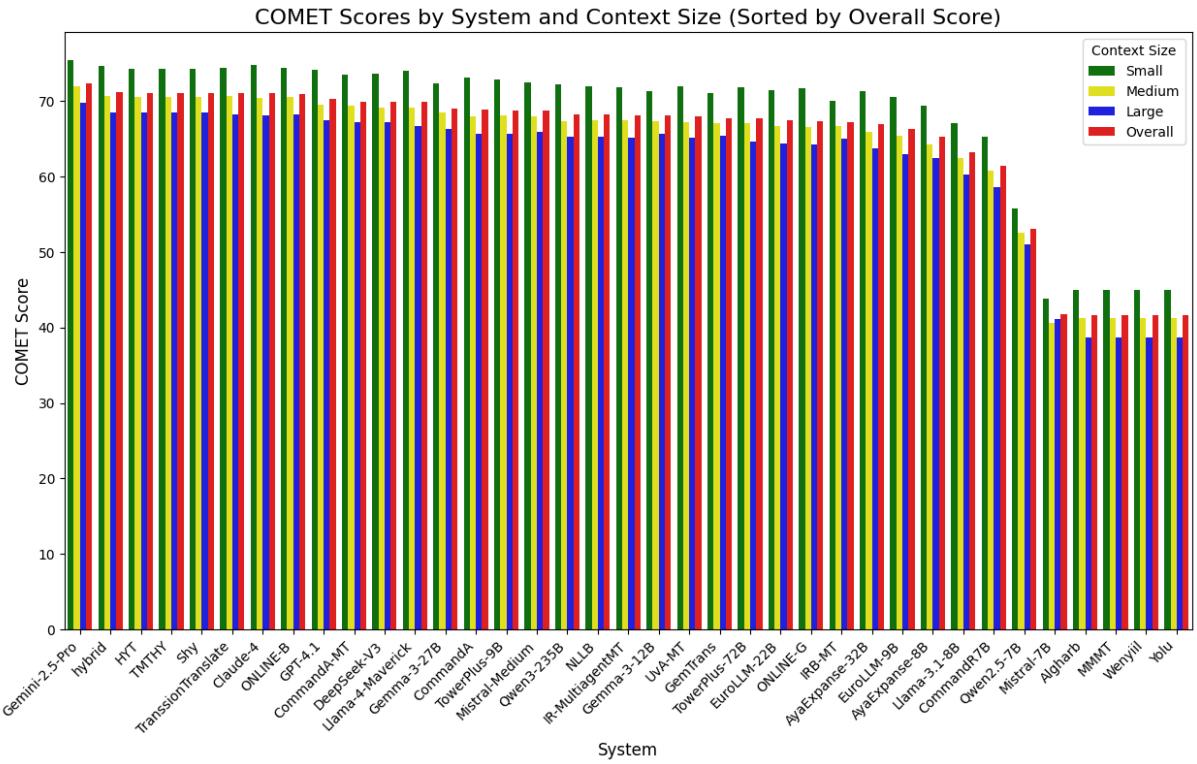


Figure 6: **COMET scores** for various translation systems, sorted by their overall performance. **Higher scores are better**. **Gemini-2.5-Pro, Hybrid, and HYT** have the best overall performance, while **MMMT, Wenyiil, and Yolu** have the lowest.

RoCS-MT v2 at WMT 2025: Robust Challenge Set for Machine Translation

Rachel Bawden **Benoît Sagot**
Inria, Paris, France
firstname.lastname@inria.fr

Abstract

RoCS-MT (Robust Challenge Set for Machine Translation) was initially proposed at the test suites track of WMT 2023. Designed to challenge MT systems' translation performance on user-generated content (UGC), it contains examples sourced from English Reddit, with manually normalised versions, aligned labelled annotation spans and reference translations in five languages. In this article, we describe version 2 of RoCS-MT in the context of the 2025 WMT test suites track. This new version contains several improvements on the initial version including (i) minor corrections of normalisation, (ii) corrections to reference translations and addition of alternative references to accommodate for different possible genders (e.g. of speakers) and (iii) a redesign and re-annotation of normalisation spans for further analysis of different non-standard UGC phenomena. We describe these changes and provide results and preliminary analysis of the MT submissions to the 2025 general translation task.

1 Introduction

Large language models (LLMs) have truly arrived in the field of Machine Translation (MT); their performance often rivals that of dedicated MT systems across various domains (Kocmi et al., 2023; Xu et al., 2024; Cui et al., 2025). However, while they have opened up new possibilities for translation, enabling fine-grained control of output formats, styles and formalities, they are also characterised by new types of errors that were less present with dedicated models, such as translating in the wrong language, inappropriate copying of the source text, generation of outputs that are not translations of the source text, etc. Evaluation continues to be a highly important aspect of the field, and as MT technologies progress, so does the way in which we evaluate. The WMT shared tasks are a good example of this, with an evolution a few years ago to general translation instead of news translation (Kocmi et al., 2022), in order to challenge systems to translate a wider range of domains. One of the selected domains is social media content, known for covering a wide range of topics and containing non-standard language typical of user-generated content (UGC) (Foster, 2010; Seddah et al., 2012; Eisenstein, 2013; Baldwin and Li, 2015; van der Goot et al., 2018).

The translation of UGC has been a topic for a number of years (Belinkov and Bisk, 2018; Michel and Neubig, 2018; Vaibhav et al., 2019; Park et al., 2020; Nishimwe et al., 2024; Peters and Martins, 2025). In particular, a shared task was organised on the matter at WMT in 2023 (Kocmi et al., 2023), designed to target non-standard language from Reddit forums. Several parallel test sets of UGC texts exist (Ling et al., 2013; Vicente et al., 2016; Sluyter-Gäthje et al., 2018; Michel and Neubig, 2018; Rosales Núñez et al., 2019; Mubarak et al., 2020; Fujii et al., 2020; McNamee and Duh, 2022), for a range of languages, although they differ according to the language pairs covered and the degree of non-standardness present.

In the 2023 edition of the test suite track at WMT, we proposed the RoCS-MT test suite (Robust Challenge Set for MT) (Bawden and Sagot, 2023), designed to contain particularly challenging sentences with respect to their non-standard nature (e.g. with spelling mistakes, use of acronyms, marks of expressiveness, devowelling, contractions, etc.). Sourced from English Reddit, we aimed to cover a range of these phenomena in the texts selected, which we manually segmented into sentences, normalised into standard English (according to guidelines that normalised as reasonably possible whilst preserving fluency and meaning) and then translated by professional translators into French, German, Czech, Ukrainian and Russian.

For this 2025 edition of WMT (Kocmi et al., 2025), we resubmit RoCS-MT in an improved version (v2), after (i) some minor corrections to the

Proceedings of the Tenth Conference on Machine Translation, pages 834–849
November 8-9, 2025

source-side normalisations, (ii) corrections to the existing references and addition of multiple references to account for different genders and (iii) improvements to the annotations of the non-standard phenomena for additional analysis. We release this version in Huggingface's Datasets (Lhoest et al., 2021).[1] In this paper we describe those changes and provide results and analysis for the WMT2025 MT systems, with a major difference being that all systems this year were applied at the document level (at the level of Reddit post text in out case). We compare different segmentations of the texts and the performance of systems when applied to the original and normalised source texts.

2 The Test Suite

Composition The main composition of the challenge set is described in the article presenting the first version (Bawden and Sagot, 2023). The English source texts are taken from Reddit (all varieties of English including some non-native language, although we avoided code-switching). Candidate posts were identified using keyword searches on the Reddit API and chunks of text were manually selected from within those posts. The selected texts were manually segmented into sentences (non-trivial since many texts did not contain standard punctuation and sometimes contained newlines within sentences) and manually normalised. The normalisation guidelines we drew up aimed to balance (i) normalising as much as possible and at the same time (ii) rendering the output natural and realistic and (iii) not over-normalising to avoid losing the original text's style. For example, we did not use normalised variants that could be spontaneously and naturally used (e.g. we kept *lol* instead of *laughing out loud*. Finally, translations of the normalised texts were professionally produced in five languages: French, German, Czech, Ukrainian and Russian. Although not all these language directions are represented in this year's shared task, these references remain relevant for future use of the challenge set for these five languages.

Changes with respect to the First Version Several changes were carried out in this second version of the test suite, namely:

- Minor corrections to the normalisations of the source-side texts; we corrected a few wrong normalisations and typos and reverted some hypercorrections to make sure that certain grammatical variations due to dialectal differences were conserved (i.e. not over-normalising)
- Corrections to some references after a manual check, including making sure that emojis and emoticons were always included in the references (this was not the case previously, notably for the Russian translations). For certain target languages, we also complete the reference translations with alternations for multiple possible genders where appropriate (namely where the gender is underspecified given the available context of the Reddit post).
- Re-annotation of the normalisation spans according to a new annotation scheme that is organised hierarchically and structures the types differently.

The new annotation types can be found below. Detailed descriptions including examples can be found in Appendix A.

- **Punctuation, typographic conventions, symbols, etc.**: punct:diff, punct:norm, caps, slash_to_or, slash_to_and, slash_distribution, word_to_symbol, symbol_placement
- **Spacing**: spacing, spacing:camelcase
- **Phonetically similar spellings (including imitation of speech)**: phon, phon:char, phon:digits, phon:cute, phon:hesitate, phon:sound, phon:interjection
- **Other spelling variations (ergographic, expressiveness)**: elongation, devowelling, contraction, truncation, acronym, abbreviation
- **Spelling mistakes**: spell, spell:charswap
- **Misc**: digit_letter_sim, letter_to_digit, suffix
- **Added and dropped words**: word_drop, word_drop:pronoun, word_drop:det, word_add, word_add:det, symbol_drop, symbol_add
- **Grammar**: inflection, grammar, grammar:v, grammar:v:inflect
- **Lexical changes**: lex_choice, surrounding_emphasis, emoticon, censored

[1] https://huggingface.co/datasets/rbawden/RoCS-MT-v2

3 WMT 2025 submissions

There were 56 submissions to the 2025 general task (including variants of the same systems and systems run by the general MT task organisers) that translated the test suite. A range of architectures were used, with a majority using LLMs. In a bid to encourage document-level translation, one of the important factors to be taken into account in MT evaluation (Läubli et al., 2018), this year's general MT task focused on document-level MT, where documents were typically paragraphs of text. For the test suites, individual segments as provided in the test suite were concatenated to form source documents to be translated in one go. RoCS-MT was provided in two different formats: (i) manual segmentation and (ii) segmentation purely based on newlines within posts.

The language directions of the 2025 shared task overlap somewhat with the 2023 language directions (e.g. English to Czech, Chinese, Japanese, Ukrainian and Russian), although not all target languages for which we have reference translations are present (e.g. French and German), and many of the language directions are new and therefore do not have references (e.g. Arabic, Bhojpuri, Estonian, Icelandic, Italian, Korean, Massai, Serbian). We therefore choose in our initial results and analysis (see Section 4) to concentrate on quality estimation (QE) (i.e. without using the references for automatic analysis). We use two different metrics to calculate the scores (at the document level) used for the rankings: (i) CometKiwi (Rei et al., 2022),[2] and (ii) MetricX (Juraska et al., 2024).[3] This is to avoid bias towards a single metric, especially as many systems optimise for a particular QE metric. We acknowledge however that (i) these metrics may well have issues handling certain languages, particularly those not explicitly included in the training data of the underlying models, and (ii) the scores may well favour models that optimise for QE in general, even if the same QE model is not used.

4 Results and Analysis

In this section, we present results for the submitted systems along with several brief analyses. In order to compute rankings, we took into account the fact that not all systems took part in every language pair, and adopted the ranking algorithm commonly used for nations in the Olympic Games, which addresses a comparable situation. We observed that if systems are ranked based on their absolute scores for each language pair, a few systems are consistently ranked first, meaning that most systems never achieve a rank of 1 or 2, even if they have relatively high overall scores. The result of this is that an overall weaker system that has low scores in most languages but happens to be ranked highly for a single language pair can be ranked higher than a system that has high but not the best scores across all language pairs. We therefore choose to apply the "gold first" algorithm on quartile-based rankings instead of raw rankings; for a given language pair, all systems within the top 10% of scores are assigned to rank 1 and treated as such by the "gold first" algorithm, the next 10% to rank 2, and so on. In other words, systems are then ordered according to the number of language pairs in which they achieve rank 1 (i.e. belong to the top 10%). In the event of a tie, the number of rank 2 placements (top 10–20%) is considered, followed, if necessary, by the number of rank 3 placements (top 20–30%), and so on.

We applied this approach to get rankings from each of the two metrics used. These rankings are based on the original inputs (i.e. before normalisation) that have been manually segmented into sentences. We then computed overall rankings based on the average of the two rankings. Table 1 displays both metric-specific rankings and the overall ranking, the number of first-, second- and third-ranks for each system and each metric and the absolute gap between each system's CometKiwi-based and MetricX-based ranks, to get an idea of how consistent they are and the confidence we can place in the resulting rankings. Appendix B provides raw CometKiwi and MetricX scores per language pair, first comparing scores on original and normalised inputs, then comparing scores on manual and newline segmentations applied to original inputs.

Results highlight a small group of systems that dominate performance. Yolu occupies the top position with nine first-place results across all twelve language pairs on both metrics, followed closely by Shy-hunyuan-MT and CommandA-WMT. Laniqo and SalamandraTA follow, despite the fact that Laniqo participated in only seven pairs. Among the organiser-run systems, GPT-4.1 ranks 7th, closely followed by ONLINE-B and both TowerPlus models. Several larger LLM-based baselines, such as Claude-4 (18th) and both Gemini models (also

[2] WMT22-COMETKIWI-DA model. Higher is better.

[3] METRICX-24-HYBRID-XL-V2P6 model. Lower is better.

System	#lp	CometKiwi		MetricX		Overall rank	Δrank
		Rank	"Medals"	Rank	"Medals"		
Yolu	12	1	9, 1, 0	3	9, 1, 1	1	2
Shy-hunyuan-MT	12	4	6, 4, 2	1	11, 0, 1	2	3
CommandA-WMT	12	2	7, 3, 1	4	8, 2, 0	3	2
Laniqo$^{\diamond}$	7	5	6, 1, 0	5	4, 3, 0	4	0
SalamandraTA	11	3	7, 2, 1	8	1, 0, 1	5	5
GemTrans	12	12	1, 2, 1	2	10, 1, 0	6	10
UvA-MT	12	7	2, 4, 4	16	0, 5, 3	7	9
*GPT-4.1	12	17	0, 4, 1	6	1, 8, 1	7	11
*ONLINE-B	11	8	2, 1, 4	20	0, 2, 2	9	12
*TowerPlus 9B	12	9	1, 5, 2	21	0, 2, 0	10	12
*TowerPlus 72B	12	11	1, 2, 3	22	0, 2, 0	11	11
SRPOL	7	6	3, 4, 0	28	0, 1, 0	12	22
IR-MultiagentMT	12	23	0, 1, 1	17	0, 4, 4	13	6
TranssionTranslate	12	10	1, 2, 4	33	0, 0, 2	14	23
*CommandA	12	18	0, 2, 4	26	0, 1, 1	15	8
NNTSU	1	32	0, 0, 1	12	1, 0, 0	15	20
Erlendur	1	32	0, 0, 1	12	1, 0, 0	15	20
In2x	1	15	1, 0, 0	30	0, 1, 0	18	15
*Claude4	12	21	0, 1, 2	24	0, 1, 2	18	3
*DeepSeek V3	12	22	0, 1, 2	23	0, 1, 7	18	1
Algharb$^{\diamond}$	12	26	0, 1, 0	19	0, 2, 3	18	7
*Gemini 2.5 Pro	12	38	0, 0, 0	7	1, 1, 7	18	31
*Gemma 3 27B	12	28	0, 0, 1	18	0, 3, 4	23	10
*AyaExpanse 32B	12	13	1, 0, 2	35	0, 0, 1	24	22
*AyaExpanse 8B	12	20	0, 2, 0	29	0, 1, 0	25	9
KIKIS	1	39	0, 0, 0	12	1, 0, 0	26	27
*EuroLLM 22B	12	19	0, 2, 0	36	0, 0, 1	27	17
*Gemma 3 12B	12	25	0, 1, 0	32	0, 0, 2	28	7
Yandex	1	46	0, 0, 0	12	1, 0, 0	29	34
Systran$^{\diamond}$	1	15	1, 0, 0	44	0, 0, 0	30	29
*Llama 3.1 8B	12	14	1, 0, 0	47	0, 0, 0	31	33
Kaze-MT$^{\diamond}$	12	52	0, 0, 0	9	1, 0, 0	31	43
KYUoM$^{\diamond}$	12	52	0, 0, 0	10	1, 0, 0	33	42
ctpc_nlp	12	52	0, 0, 0	11	1, 0, 0	34	41
Wenyiil$^{\diamond}$	12	30	0, 0, 1	34	0, 0, 2	35	4
*Mistral-Medium	9	40	0, 0, 0	25	0, 1, 1	36	15
*CommandR	12	24	0, 1, 1	43	0, 0, 0	37	19
*Qwen3 235B	12	41	0, 0, 0	27	0, 1, 1	38	14
*ONLINE-W	8	27	0, 1, 0	42	0, 0, 0	39	15
AMI$^{\diamond}$	1	32	0, 0, 1	37	0, 0, 1	39	5
*EuroLLM 9B	12	29	0, 0, 1	41	0, 0, 0	41	12
IRB-MT	12	42	0, 0, 0	31	0, 0, 3	42	11
*Llama-4-Maverick	12	37	0, 0, 0	38	0, 0, 0	43	1
CUNI-MH-v2	1	32	0, 0, 1	46	0, 0, 0	44	14
bb88	1	32	0, 0, 1	49	0, 0, 0	45	17
*NLLB	12	44	0, 0, 0	39	0, 0, 0	46	5
*Mistral 7B	12	31	0, 0, 1	53	0, 0, 0	47	22
DLUT_GTCOM	2	45	0, 0, 0	40	0, 0, 0	48	5
CUNI-SFT	3	48	0, 0, 0	45	0, 0, 0	49	3
TranssionMT	8	43	0, 0, 0	51	0, 0, 0	50	8
*Qwen 2.5	12	47	0, 0, 0	48	0, 0, 0	51	1
CGFOKUS	1	51	0, 0, 0	49	0, 0, 0	52	2
*ONLINE-G	10	49	0, 0, 0	52	0, 0, 0	53	3
SH	1	50	0, 0, 0	54	0, 0, 0	54	4
CUNI-DocTransformer	1	55	0, 0, 0	55	0, 0, 0	55	0
COILD-BHO	1	55	0, 0, 0	55	0, 0, 0	55	0

Table 1: Main ranking table. For each system and for both the CometKiwi and MetricX metrics, we provide their rank according to the metric, computed using the "gold first" scoring algorithm (Rank), the number of language pairs for which the system ranked first, second and third ("Medals"). For each system we also provide the number of language pairs it participated in (#lp), an overall rank based on the average between the CometKiwi- and the MetricX-based ranks, and the difference between the two original ranks, which shows for each system how consistent the two metrics are. Systems run by the task organisers are marked with an asterisk, while systems fine-tuned to optimise a Qe metric, such as CometKiwi and MetricX, are indicated with a $^{\diamond}$.

18th and 23rd, respectively), appear in the middle of the table, while Llama-4-Maverick, Mistral 7B, Qwen 2.5 and ONLINE-G fall into the lower ranks. NLLB, a widely used dedicated MT model, ranks only 46th, with consistently low ranks in both metric-specific rankings (44th and 39th). Conversely, several single-pair submissions are quite successful, achieving first place according to one of the metrics, and sometimes second or third according to the other one (NNTSU, Erelendur and In2x reach a joint overall 15th place). Overall, the results indicate that dedicated MT systems continue to outperform many general-purpose LLMs when evaluated using CometKiwi and MetricX.

These results are somewhat surprising, particularly the relatively low performance of several large LLMs—Qwen3 235B ranks only 38th overall, and GPT-4.1 is ranked 17th according to the CometKiwi-based ranking—and popular reference models such as NLLB, despite generally being evaluated quite highly in MT tasks, whether by automatic metrics or human evaluation. Three main factors may account for this outcome. First, strong performance on edited data does not necessarily translate into equally strong performance on non-standard data; success on the former is not a guarantee of robustness. Secondly, automatic evaluation metrics—especially CometKiwi, on which our ranking is based—may perform poorly when applied to translations of non-standard text. Thirdly, several systems used QE metrics for optimisation and may therefore have gained higher rankings than their actual quality would otherwise justify.[4] We leave these questions open for future investigation, and invite the reader to take our results and conclusions with a pinch of salt.

Another surprising observation is that several systems are positioned very differently in our two metric-specific rankings. The most extreme case is KazeMT, ranked 52nd using CometKiwi but 9th using MetricX. Another example of a large Δrank is Gemini 2.5 Pro, ranked 38th using CometKiwi but 7th using MetricX. Several single-pair submissions also display large discrepancies, such as Yandex and Systran, which both achieve first place according to one metric (MetricX for Yandex and CometKiwi for Systran), but do not perform as well according to the other metric. Such discrepancies could be explained, at least in some cases, in the way these models were trained or fine-tuned, for instance by optimising for QE, as mentioned above.

[4]We indicate systems that self-declare as using QE in some way with a $^{\diamond}$.

4.1 Original versus Normalised Texts

We first compare the impact of translating the original inputs (containing non-standard language) against the normalised inputs (both with manual segmentation). The full results are given in Appendix B (Tables 3 and 4). The scores for original texts are in general lower for CometKiwi and higher for MetricX than for the normalised ones. This is somewhat unsurprising for several reasons: (i) it is expected that more standard texts are easier to translate, as the majority of the texts that the models were trained on was standard, and the UGC texts are characterised by high levels of variation, (ii) metric scores are likely to penalise translations that are less standard. Concerning (i), there is some indication that is going on. For example, the difference between translations from normalised and original texts is very large for the lowest-resource language directions, at least for CometKiwi (English to Icelandic and to Maasai), showing that the models are struggling more with the non-standard texts. Concerning (ii), some further investigation is necessary here to ascertain whether the difference in scores are a property of the metrics themselves or whether they translate into real differences in translation quality. Our observations in Section 4.3 indicate that there is more going on than these basic scores and that we should not trust the metric scores alone.

4.2 Manual Segmentation versus Newline Segmentation

We then compare the impact of the text segmentation by looking at the scores based on inputs with manual segmentation and those separated on newlines (both with raw inputs). Results are shown in the same appendix section (Tables 5 and 6). The differences between the two segmentation types appears less than the differences previously observed between original and normalised inputs. In reality, given that the posts were given as complete documents, the segmentation has less of an impact than if the systems had been translating on the sentence level, as was the case for most systems in previous years.

4.3 A First Qualitative discussion

Table 2 shows the results of all systems (apart from those whose output was obviously the result of a

System	Text	CometKiwi	MetricX	tgt lang	elong tr
Source	*"politics used to be moooorrreeeee poollliitte we neeeeed a return to The Discouuurrrse"*				
Qwen 2.5	politika byla dávno většinou politická, potřebujeme návrat k The Discouuurrrse.	45.4	17.4	+	−−−
Llama 3.1 8B	politika byla kdysi mnohem menší, potřebujeme návrat k diskuzi.	46.2	15.9	+++	−−−
ctpc_nlp	„politics used to be moooorrreeee poollliitte we neeeeed a return to The Discouuurrrse.“	46.3	17.4	−−−	−−−
Mistral 7B	politika byla dávno více politická, potřebujeme se vrátil do The Discouuurrrse.	47.7	16.9	+	−−−
GPT-4.1	„politika býýýýýala mnohem slušnějššííí, potřebujeme návrat k diskuuuziii.“	49.6	14.6	+++	++
CUNI-SFT	politika bývala moooorrreeeee poollliitte, potřebujeme návrat k diskusi.	49.6	16.3	–	−−−
Gemma 3 27B	„politika bývala mooooc víc slušná, potřebujeme návrat k Diissskuuzi.“	51.3	10.9	+++	++
EuroLLM 9B	„politika bývala moooorrreeeee poollliitte, potřebujeme návrat k Diskouuurrrse.“	51.4	18.0	−−	+
CUNI-DocTransformer	„politika bývala moooorrreeeee poollliitte my jsme potřebovali návrat do The Discourrrse.“	51.8	17.9	−−	−−−
Gemini 2.5 Pro	„politika bejvávala slu-šně-ě-ě-jšíííí, po-tře-bu-je-meee se vrátit k Diskurzuuuu.“	52.8	11.5	++	+!
ONLINE-W	politics used to be moooorrreeeee poollliitte we neeeeed a return to The Discouuurrrse.	52.8	17.7	−−−	−−−
Wenyiil	„politika bejvááaaála kdysi slušněěěě, potřebíííijem návrat k Diskouuuurrrsu“	53.6	10.7	++(!)	+++
CUNI-MH-v2	politika bývala moooorrreeeee poollliitte, potřebujeme návrat k Diskurzu.	53.6	18.4	–	−−−
Algharb	„politika bejvááááála kdysi slušněěěěějšššší, musíme se vrááátit k Diiiskuuurrrzu“	53.7	10.0	+++(!)	+++
AyaExpanse-32B	politika bývala víceeee čistáá, potřebujeme se vrátit k Diiskusi.	54.5	14.0	+++	++
SRPOL	„politika bývala moooorrreeeee poollliitte, potřebujeme návrat k diskuusii.“	54.6	16.2	–	+
GemTrans	„politika bývala kdysi mnohem civilnější a potřebujeme návrat k seriózní debatě.“	54.7	4.6	+++	−−−
Tower Plus 72B	„politika bývala více politická, potřebujeme návrat k Diskurzu.“	55.1	10.8	+++	−−−
IRB-MT	politika bývala dřív mnohem uhlazenější, potřebujeme návrat k slušné konverzaci.	55.1	7.8	+++	−−−
Qwen3 235B	„dřív byla politika víc vstřícná, potřebujeme se vrátit ke civilizované debatě.“	55.3	9.6	+++	−−−
Gemma 3 12B	politika bývala dřív mnohem, mnohem političtější, potřebujeme návrat k Diskuuzi.	55.8	10.8	+++	−−−
CommandA	„politika bývala mooоc poolliittická, potřebujeme návrat k Diissccouurrrse.“	56.2	12.2	–	+
Yolu	politika dřív bývala mnohem... mnohem... polemickyjší a my potřebujeme návrat k společné debatě.	56.3	9.0	+++	+(!)
Shy-hunyuan-MT	„politika dřív byla mnohem slušnější, potřebujeme návrat k racionálnímu diskurzu“	57.3	7.6	+++	−−−
Claude-4	politika bývala víiíc sluuušnááá, potřebujeme se vrátit k Diskuuuurzu.	57.5	12.1	+++	+
DeepSeek-V3	„politika bývala slušnějšíííí, musíme se vrátit k diskurzuuuu“	57.5	9.8	+++	+
IR-MultiagentMT	politika bývala mnohem více otevřená, potřebujeme se vrátit k diskurzu.	57.7	10.1	+++	−−−
Mistral-Medium	politika bývala kdysiiiii moooooc civilizovanější, my potřebujeme návrat k Diskuuuuuursu.	58.0	9.0	+++	++
Laniqo	„politika se stala blbá a potřebujeme návrat ke klasickému politickému diskurzu.“	58.3	9.0	+++	−−−
UvA-MT	politika bývala mnohem političtější, potřebujeme návrat k Diskuuzi.	58.5	8.9	+++	+
TowerPlus 9B	„politika bývala mooooоc lepší, potřebujeme návrat k debatám.“	58.7	9.7	+++	+
Llama-4-Maverick	politika bývala mnohem uhlazenější, potřebujeme návrat k onomu Diskurzu.	58.7	10.1	+++	−−−
AyaExpanse 8B	politika byla kdysi zábavnější, potřebujeme se vrátit k diskusi.	60.1	12.1	+++	−−−
CommandA-WMT	„politika bývala víc politická, potřebujeme se vrátit k diskurzu“	61.9	6.8	+++	−−−
SalamandraTA	„politika bývala mnohem menší, takže jsme potřebovali návrat k The Discouuurrrse“	62.5	11.8	++	−−−
TranssionTranslate	„politika bývala moooooooočná, že potřebujeme návrat k The Discouurrse.“	64.8	21.6	++	+
ONLINE-B	„politika bývala moooooodně dobrá, potřebujeme návrat k The Discouurrse.“	66.7	11.0	++	+
NLLB	Politika bývala moooorrreeeee poollliitte, potřebujeme návrat do The Discouuurrrse	68.0	15.3	−−	−−−

Table 2: Example of character repetition linked to a mark of expressivity for en–cs (same source text as in (Bawden and Sagot, 2023) to illustrate 2023 en–de results). For each system we provide the CometKiwi score (multiplied by 100; higher is better) and the MetricX score (lower is better) for the corresponding document, as scores were computed at the document level. Systems are ordered by increasing CometKiwi score. The two last columns provide a manual assessment of how much of the input sentence was translated into Czech—or at least not kept in English—(“tgt lang”) and of how well the elongation phenomenon was transferred to the output sentence (“elong tr”; non-translated tokens are ignored). Systems whose outputs obviously result from an error are not included.

bug) on the example already used in (Bawden and Sagot, 2023) to illustrate the behaviour of MT systems in the presence of several instances of the elongation phenomenon, by which one or more characters are repeated to express emphasis. A first glance at the results shows that there is not necessarily a convincing correlation between perceived translation quality and the automatic evaluation provided by the CometKiwi metric, whereas MetricX results look slightly more correlated. Looking more closely at the translations, two main observations can be made:

- Firstly, a number of systems tend to keep unchanged original English tokens that have undergone elongation, and sometimes even the whole input. The fact that the two last tokens are capitalised in the input sentence makes it even more difficult for most systems to actually try to translate them.
- Secondly, not all systems attempt to transfer the elongation phenomenon into their output. Some seem to (try to) produce standard Czech rather than preserving the non standard expressivity mark. Some even try to render the same expressivity using another non standard phenomenon.

To better understand what is at play here, we decided to manually annotate these translations for two features: how much of the input sentence was (tentatively) translated into Czech, and how much of the elongation phenomenon was transferred into the output sentence (ignoring tokens kept in their original English form). Comparing these annotations with system types is interesting. Although a single example is in no way sufficient to allow for any generalisations, it seems that generic LLMs are more liable to preserving elongation and, more generally, to produce better translations, whereas

dedicated MT models seem to produce more standard outputs and/or not to translate significant parts of the input. Interestingly, this is not reflected in the CometKiwi scores, but it is more visible in the MetricX scores. For instance, the best CometKiwi-scored translation contains two segments that are still in English, a situation that invariably leads to bad (high) MetricX scores. However, CometKiwi ad MetricX seem consistently bad at penalising the absence of elongations in the produced translation. The best CometKiwi-scored translation does not contain any elongation in genuinely Czech tokens, and the best Metric-X-scored translation, which is perfect Czech, does not include any elongation whatsoever. On the contrary, the output of GPT-4.1 is good in both regards—it is entirely in Czech and does contain elongations—, and is a good translation, but it is scored poorly by both CometKiwi and MetricX. This shows that modern metrics such as CometKiwi and MetricX might not be reliable when it comes to assessing translation quality of non-standard content. We leave a more quantitative and systematic exploration of these questions and their implications for MT evaluation in general to future work.

Although the test suite this year presents new annotations for the non-standard phenomena present in the test suite that are more consistent and interesting for analysis, we also leave the analysis on a per-phenomenon basis to future work, in which we will go into more detail and length.

5 Conclusion

We have presented a new version (v2) of the RoCS-MT challenge set, first presented at the WMT 2023 test suites shared task track. This 2025 edition has several improvements, with minor corrections to source texts, some corrections to references and improved categorisation of non-standard phenomena. We describe these changes and also use the challenge set to compare systems submitted to this year's shared task, comparing translation from the original UGC inputs and their manually normalised versions. A major difference with previous years of the shared task is a switch to document-level MT, so whole chunks of posts were submitted to systems for translation. We nevertheless compare two different segmentation types (to see if initially manually segmenting into sentences and then concatenating the sentences with newlines could help translation) and discuss preliminary insights into the shortcomings of popular metrics such as CometKiwi and MetricX when applied on non-standard text MT.

Acknowledgments

This work was partly funded by both authors' chairs in the PRAIRIE institute, now PRAIRIE-PSAI, funded by the French national agency ANR, respectively as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and as part of the "France 2030" strategy under the reference ANR-23-IACL-0008.

References

Tyler Baldwin and Yunyao Li. 2015. An in-depth analysis of the effect of text normalization in social media. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 420–429, Denver, Colorado. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness challenge set for machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 198–216, Singapore. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 359–369, Atlanta, Georgia. Association for Computational Linguistics.

Jennifer Foster. 2010. "cba to check the spelling": Investigating Parser Performance on Discussion Forum Posts. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 381–384, Los Angeles, California. Association for Computational Linguistics.

Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui.

2020. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5929–5943, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria. Association for Computational Linguistics.

Paul McNamee and Kevin Duh. 2022. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918, Marseille, France. European Language Resources Association.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2020. Constructing a bilingual corpus of parallel tweets. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 14–21, Marseille, France. European Language Resources Association.

Lydia Nishimwe, Benoît Sagot, and Rachel Bawden. 2024. Making sentence embeddings robust to user-generated content. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10984–10998, Torino, Italia. ELRA and ICCL.

Jungsoo Park, Mujeen Sung, Jinhyuk Lee, and Jaewoo Kang. 2020. Adversarial subword regularization for robust neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1945–1953, Online. Association for Computational Linguistics.

Ben Peters and Andre Martins. 2025. Did translation models get more robust without anyone Even noticing? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*,

pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT Performance for Translating Noisy User-Generated Content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a Treebank of Noisy User Generated Content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Henny Sluyter-Gäthje, Pintu Lohar, Haithem Afli, and Andy Way. 2018. FooTweets: A bilingual parallel corpus of world cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vaibhav Vaibhav, Sumeet Singh, Craig Stewart, and Graham Neubig. 2019. Improving robustness of machine translation with synthetic noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1916–1920, Minneapolis, Minnesota. Association for Computational Linguistics.

Rob van der Goot, Rik van Noord, and Gertjan van Noord. 2018. A taxonomy for in-depth evaluation of normalization for user generated content. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 684–688, Miyazaki, Japan. European Language Resources Association (ELRA).

Iñaki San Vicente, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. TweetMT: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2936–2941, Portorož, Slovenia. European Language Resources Association (ELRA).

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive Preference Optimization: Pushing the Boundaries of LLM Performance in Machine Translation. In *Proceedings of the Forty-First International Conference on Machine Learning*.

A Normalisation Span Classification

The original and normalised texts were aligned and different non-standard phenomena classified. Below is the list of normalisation categories with examples.

Punctuation, typographic conventions, symbols, etc.

- **punct:diff**: extra punctuation is included or necessary punctuation is removed (e.g. missing final punctuation, missing apostrophes, commas, etc.).
  e.g. im→I'm
- **punct:norm**: punctuation to be normalised according to certain conventions (e.g. same apostrophes and quotes).
  e.g. that's→that's
- **caps**: capitalisation differs from what is considered standard (e.g. lowercase initial characters, all uppercase, etc.).
  e.g. IM SO HAPPY→I'm so happy
- **slash_to_or**: a slash is used, where in normalised speech an "or" would be used to represent a list of items. This applies to the whole list, including where etc. is included. Not that this does not include cases where the items are alternatives in the discourse
  e.g. cat/exhaust/etc→cat or exhaust, etc.
  e.g. truth/dare→truth or dare
  e.g. [counter-example] AW WELL MY DOG/CHILD IS VERY FRIENDLY SO LET ME APPROACH→aw, well my dog/child is very friendly, so let me approach
- **slash_to_and**: a slash is used, where in normalised speech an "and" would be used. This applies to the whole list, including where etc. is included.
  e.g. work/paint→work and paint
- **slash_distribution**: the use of a slash to separate two items where the slash does not separate two complete items (i.e. part of one element is distributed to both items thanks to the slash). An example makes this easier to understand:
  e.g. just disrespects any / everyone→just disrespects any / everyone
- word_to_symbol: the use of a symbol to represent a word
  e.g. +→and, &→and
  e.g. →around
  e.g. $$$→money

- **symbol_placement**: non-standard placement of a symbol with respects to English norms.
  e.g. 100$→$100

Spacing

- **spacing:** missing or added spacing in the original text
  e.g. aswell→as well
  e.g. over thinking→overthinking

- **spacing:camelcase:** the use camelcase (capital letters at the beginning of words) instead of using spaces
  e.g. surroundedUs→surrounded us
  e.g. sawThat→saw that

Phonetically similar spellings (including imitation of speech)

- **phon:** the word uses a variant of spelling based on the phonetic similarity of the sequence of characters. This also includes the use of individual letters to represent words or syllables because of an equivalence in their pronunciation (u→you, b→be, c→see).
  e.g. saturday sesh→Saturday session (also a case of truncation)
  e.g. sup→What's up (also truncation)
  e.g. bcos -> because
  e.g. n→and
  e.g. tho→though
  e.g. speakin→speaking

- **phon:char:** a character is used in the place of a word or syllable because of its phonetic similarity with the word or syllable
  e.g. b→be
  e.g. c→see
  e.g. u→you

- **phon:digits:** a digit is used in the place of a word or part of a word.
  e.g. m8→mate
  e.g. 2→to
  e.g. as1 that will play→as one that will play (in a context where 1 could be incorrect, otherwise this should not be normalised)

- **phon:cute:** the spelling of a word to indicate "cute" or babyish pronunciation, e.g. using 'w' to replace an initial letter
  e.g. wecommended→recommended

- **phon:hesitate:** words that are written in a way to imitate hesitation
  e.g. terribl-....yyy→terribly
  e.g. Y-y-yyy-yes→Yes

- **phon:sound:** the case of words that are used to indicate a sound (very rare)
  e.g. bRRrrRRrrRRrr→brr rrr rrr

- **phon:interjection:** interjections that are normalised to single (and more standard) variations
  e.g. bla→blah
  e.g. URGHHH→ugh
  e.g. Nawh→no

Other spelling variations (ergographic, expressiveness)

- **elongation:** characters are repeated, usually as a a mark of expressiveness.
  e.g. *meeeeelltiiiinggg*→melting
  e.g. sooo→so

- **devowelling:** a word with the vowels removed (initial vowels are often kept however). This can often result in double characters being reduced to single ones (messages→msgs) In this category are also words where part of the word has been devowelled.
  e.g. wt→what, ovr→over, ppl→people
  e.g. askd→asked (initial vowel kept)
  e.g. travllr→traveller

- **contraction:** when several words are contracted into a single one. This has some overlap with the characteristics of phonetic distance, in that it is due to the pronunciation of the words that the contraction occurs.
  e.g. gonna→going to
  e.g. innit→isn't it?

- **truncation:** a word is shortened, either at the end (traditional truncation) or sometimes at the beginning, often by removing a syllable or a suffix. Note the difference with acronymisation, which involves keeping initial characters.
  e.g. sesh→session
  e.g. cuz→because, till→until
  e.g. ofc→of course -> CHANGED, NOW ACRONYM
  e.g. w→with -> CHANGED, NOW ACRONYM

- **acronym:** a word or sequence of word is represented as an acronym, i.e. the initial characters of the word (or syllables) are retained and the others are elided.
  e.g. RN→right now
  e.g. gf→girlfriend
  e.g. never mind→nvm
  e.g. w→with
  We also include in this category words that are partially acronymised (i.e. where one syllable is represented by its initial but the rest is not).is acronymised but the rest is not.
  e.g. oline→offensive line
  e.g. gmeet -> Google meet
  e.g. bday→Birthday
  e.g. ofc→of course
  Note that sometimes slashes are included in the acronym
  e.g. b/c→because,
  e.g. w/o→without.

- **abbreviation:** abbreviations for units of measurement and other standard cases
  e.g. ft→feet, 2k→2000, hrs→hours
  e.g. Ex→for example

Spelling mistakes (distinguished from spelling variation identified as being intentional)

- **spell:** the word contains a spelling error that is not clearly intentional (covered by the other phenomena such as truncation, devowelling, etc.) and not covered by the other more specific categories.

- **spell:charswap:** the characters in the word are present but not in the right order (most often consecutive characters being swapped)
  e.g. nobel→noble
  e.g. furhter→futher

Misc

- **digit_letter_sim:** Very rare, but where a digit is used in the place of a letter due to the typographic similarity (see in 3ver→ever).

- **letter_to_digit:** Very rare, but where a digit is used in place of a letter not because of their typographic similarity, but because as a sort of tautology (seen in 1nce→Once).

- **suffix:** the addition of a suffix to a word, either as a diminutive or other
  e.g. lolsky→lol
  e.g. meanie→mean
  e.g. doggy→dog

Added and dropped words

- **word_drop:** a word is not present in the original text and is present in the normalised version
  e.g. It also confusing. . . →It's also confusing
  e.g. u wanna see?→Do you want to see?

- **word_drop:pronoun:** the original text omits a pronoun (often the case of subject pronouns at the beginning of sentences) that is included in the normalised version.
  e.g. Was gunna try distortion. . . →I was going to try distortion. . .

- **word_drop:det:** the original text omits an article (e.g. the or a) that is included in the normalised version.
  e.g. Pretty creative way. . . →A pretty create way. . . word_add: a word is present in the original text and is removed in the normalised version
  e.g. . . . in ten days ago→. . . ten days ago
  e.g. also for uses of word "like" as a filler

- **word_add:det:** the original text includes an article where the normalised version removes it
  e.g. . . . adds an 12kg of salt→. . . adds 12kg of salt symbol_drop: the original text omits a symbol that is included in the normalised version.
  e.g. 32c→32°C

- **symbol_add:** the original text includes a symbol that is removed in the normalised version.
  . . . no issue w being over 12+ ft. . . →. . . no issue with being over 12 feet. . .

Grammar

- **inflection:** a word is not correctly inflected (e.g. with respect to number, tense, etc.)
  e.g. . . . wondering what ppl thought are→. . . wondering what people's thoughts are

- **grammar:** inflection-related errors
  e.g. . . . wondering what ppl thought are→. . . wondering what people's thoughts are
  e.g. if your good→if you're good

- **grammar:v:**
- **grammar:v:inflect**

Lexical changes

- **lex_choice:** a use of a non-standard lexical choice, including dialectisms (e.g. cannae, ain't), malapropisms (e.g. genually), foreign words and generally wrong choices of words (e.g. wrong part of speech, wrong semantic choice of words, lacking punctuation, use of an antonym by accident, etc.)
  e.g. I am confusion→I am confused
  e.g. genually→genuinely
  e.g. pish→piss
  e.g. ain't→aren't
  e.g. cannae→cannot
  e.g. y'all→everyone/all/all your (depending on the context)
  e.g. sans guac→without guacamole

- **surrounding_emphasis:** emphasis added to certain words typographically (removed in the normalised variants).
  e.g. *without*→without
  e.g. ∼find∼→find

- emoticon: emoticon that is a variant on the common emoticons :-), :-D, :-(, :-/ and >:-)
  e.g. :-/////→:-/
  e.g. (:→:-)
  e.g. :$\hat{)}$→:-)

- censored: the word contains symbols in an effort to censor the word
  e.g. upv*te→upvote
  e.g. s**t→shit, sh**→shit

B Raw Automatic Scores per Language pair

The raw scores (calculated at the document level) can be found in this appendix section. In each of the tables, the systems are ordered by the ranking across all languages for that particular metric (as described in Section 4). Note that higher CometKiwi scores are better and lower MetricX scores are better.

Tables 3 and 6 provide the CometKiwi and Metricx scores respectively for manually segmented texts, with a comparison of original and normalised input texts.

Tables 5 and 6 provide the CometKiwi and Metricx scores respectively for original inputs, with a comparison of manually segmented and newline-segmented texts.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig
Yolu	1	78.5	73.4	71.9	68.9	82.1	76.3	84.8	79.6	35.9	22.7	83.7	79.7	83.3	79.3	35.9	22.7	81.1	76.0	83.8	78.5	80.7	75.5	80.7	76.1
CommandA-WMT	2	71.4	65.6	75.5	72.8	80.9	75.1	83.0	77.8	75.0	70.0	82.7	78.5	82.5	78.1	65.5	62.6	80.0	74.7	80.6	75.0	80.0	75.3	79.6	74.7
SalamandraTA	3	72.1	66.8	60.7	57.4	81.9	76.0	84.7	78.9	77.6	72.2	82.1	78.4	81.1	77.1	–	–	81.1	75.5	83.7	78.3	80.5	75.1	80.1	75.4
Shy-hunyuan-MT	4	75.8	70.7	80.5	76.6	80.6	74.8	83.2	77.6	77.5	71.4	82.1	77.7	81.5	76.9	55.9	52.2	79.4	73.9	82.8	77.0	79.3	73.7	79.2	73.9
Laniqo	5	–	–	–	–	80.8	74.7	83.9	78.0	–	–	82.4	78.3	82.2	78.1	–	–	80.3	74.7	–	–	79.9	74.6	79.7	74.7
SRPOL	6	76.8	71.2	–	–	81.0	74.4	83.1	77.4	–	–	82.5	78.5	–	–	–	–	79.7	74.1	–	–	79.3	73.8	79.3	74.2
UvA-MT	7	71.6	66.4	73.0	68.2	79.9	73.8	81.3	75.7	71.9	67.9	82.3	78.2	82.1	77.6	37.3	37.1	79.5	73.8	81.9	76.7	78.5	73.5	79.2	74.4
*ONLINE-B	8	76.4	71.4	60.0	54.5	79.5	73.1	82.1	75.8	76.7	71.4	82.3	78.3	81.2	76.6	–	–	78.7	73.3	80.3	74.5	78.3	72.1	79.1	73.8
*TowerPlus 9B	9	48.9	46.0	79.1	75.8	79.6	72.9	62.3	57.9	76.2	70.8	81.7	77.5	81.6	77.2	43.6	41.3	79.0	73.7	63.8	60.2	78.3	72.8	78.9	73.6
TranssionTranslate	10	67.5	61.7	59.9	54.9	79.5	72.9	82.2	76.0	76.5	71.1	82.8	78.5	81.9	77.1	35.9	22.7	79.4	73.4	76.6	65.3	78.1	71.7	79.1	73.8
*TowerPlus 72B	11	66.4	60.4	79.7	76.2	78.8	72.7	72.6	67.4	75.5	70.4	81.6	77.1	81.4	77.0	44.0	38.7	79.1	73.4	73.6	68.8	78.2	72.5	79.0	74.1
GemTrans	12	75.7	70.3	80.4	76.9	78.9	72.3	81.3	74.8	72.7	67.5	80.7	76.2	80.4	75.7	36.5	34.3	78.3	72.4	81.2	75.1	78.0	72.7	78.0	73.1
*AyaExpanse 32B	13	66.0	60.5	65.2	62.8	78.2	72.7	54.1	50.9	50.9	47.8	81.6	77.0	81.3	76.4	46.5	43.2	78.2	73.1	72.9	68.3	77.6	72.4	77.9	72.6
*Llama 3.1 8B	14	57.2	52.0	66.3	62.1	72.9	66.7	67.2	62.6	56.4	53.4	75.7	72.2	76.8	72.0	50.9	48.8	74.8	69.2	73.6	68.1	73.3	68.2	76.4	70.9
Systran	15	–	–	–	–	–	–	–	–	–	–	84.3	81.0	–	–	–	–	–	–	–	–	–	–	–	–
In2x	15	–	–	–	–	–	–	–	–	–	–	82.8	78.7	–	–	–	–	–	–	–	–	–	–	–	–
*GPT-4.1	17	62.7	57.9	62.4	59.6	79.8	73.2	82.5	76.5	76.2	70.5	81.6	77.2	81.5	76.8	34.6	32.2	78.6	72.6	82.0	76.2	78.2	72.5	78.5	73.1
*CommandA	18	64.1	59.3	63.0	60.2	79.5	73.5	77.3	71.3	67.9	63.7	82.1	77.7	81.9	77.4	42.1	39.9	78.7	73.1	79.5	74.5	78.2	72.5	78.9	73.8
*EuroLLM 22B	19	72.1	66.0	77.2	73.6	79.0	72.5	81.9	76.1	45.7	43.1	81.3	76.7	81.2	75.9	39.7	36.7	78.5	73.0	79.3	74.2	77.7	72.1	78.3	73.1
*AyaExpanse 8B	20	74.6	69.7	74.7	71.1	77.7	71.4	39.3	37.5	39.6	37.4	81.0	76.7	80.9	76.2	39.9	38.2	77.6	72.0	60.7	56.8	77.4	71.9	77.6	72.5
*Claude4	21	64.8	59.4	63.1	59.5	79.7	72.8	81.6	74.9	75.9	69.4	82.2	77.5	81.7	76.9	39.8	37.6	78.7	72.3	81.1	75.3	78.2	72.2	78.6	73.0
*DeepSeek V3	22	64.6	58.9	65.1	62.5	78.9	72.4	81.5	75.3	74.5	68.8	81.2	76.4	80.8	75.7	38.5	35.4	77.8	72.2	81.0	75.4	77.0	71.7	76.5	70.5
IR-MultiagentMT	23	68.2	63.7	63.4	61.5	78.3	72.5	80.7	74.9	75.0	69.6	81.3	76.4	80.7	76.5	38.4	36.2	78.0	73.0	80.9	75.2	77.7	72.5	77.9	72.9
*CommandR	24	72.8	66.5	66.0	62.2	74.2	67.8	42.4	39.5	41.1	38.7	79.9	75.3	79.2	74.6	44.6	41.2	72.4	66.1	61.6	57.2	72.5	66.9	76.7	71.4
*Gemma 3 12B	25	62.3	57.5	59.3	57.2	77.9	71.5	78.9	72.7	69.2	64.3	80.3	75.5	79.5	74.8	43.9	42.1	77.7	72.6	79.6	74.0	77.4	72.4	77.6	72.5
Algharb	26	74.8	69.2	59.1	56.7	78.1	71.6	81.4	74.9	35.9	22.7	80.5	75.8	80.0	75.1	35.9	22.7	76.9	70.6	80.9	74.7	76.8	70.9	77.2	71.3
*ONLINE-W	27	74.6	68.7	–	–	79.1	68.9	80.5	68.0	–	–	79.6	73.8	81.4	75.4	–	–	79.0	72.5	–	–	78.1	72.1	78.3	73.0
*Gemma 3 27B	28	63.9	59.1	62.4	59.8	78.5	72.5	80.9	75.1	73.2	68.1	81.0	76.5	80.2	75.4	35.3	32.6	78.1	72.4	80.7	75.0	77.7	72.5	77.8	72.5
*EuroLLM 9B	29	70.3	64.7	69.9	65.4	78.7	72.4	81.2	75.0	43.2	41.5	80.3	75.9	80.3	75.6	32.5	28.5	77.8	72.3	77.6	70.6	77.2	71.6	77.4	72.0
Wenyiil	30	74.0	67.6	58.2	55.6	76.8	68.7	79.8	72.4	35.9	22.7	79.4	74.3	78.9	73.3	35.9	22.7	76.0	69.0	79.9	72.8	76.2	69.8	76.6	70.4
*Mistral 7B	31	48.6	45.0	55.6	52.8	64.7	59.4	41.5	39.1	42.5	40.4	71.2	67.9	71.5	67.6	41.2	40.1	71.2	66.2	68.9	64.6	70.1	65.4	71.2	65.9
CUNI-MH-v2	32	–	–	–	–	79.2	73.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
NNTSU	32	–	–	–	–	–	–	–	–	–	–	82.3	78.2	–	–	–	–	–	–	–	–	–	–	–	–
bb88	32	–	–	–	–	–	–	–	–	–	–	82.1	77.8	–	–	–	–	–	–	–	–	–	–	–	–
AMI	32	–	–	–	–	–	–	–	–	75.4	69.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	32	–	–	–	–	–	–	–	–	75.0	69.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	37	62.8	57.7	60.2	57.3	78.5	71.8	81.3	74.4	72.9	67.0	80.4	75.8	80.1	75.4	37.2	34.8	78.4	72.6	80.6	74.7	78.0	72.3	78.1	72.4
*Gemini 2.5 Pro	38	58.9	54.8	59.4	56.5	77.8	70.7	80.7	74.4	74.6	68.8	80.1	75.1	79.3	74.0	37.6	35.2	76.3	69.7	80.4	74.3	76.1	70.2	76.4	70.2
KIKIS	39	–	–	–	–	–	–	–	–	–	–	81.9	77.3	–	–	–	–	–	–	–	–	–	–	–	–
*Mistral-Medium	40	65.6	60.3	63.3	60.6	79.2	72.5	79.9	73.3	72.2	66.9	81.6	76.7	81.2	76.2	–	–	–	–	–	–	78.0	72.0	78.3	72.8
*Qwen3 235B	41	64.7	59.0	62.8	60.4	77.7	71.3	76.2	69.7	68.4	63.5	81.2	76.9	80.4	76.0	38.9	36.4	78.3	72.2	78.6	72.5	76.8	71.2	78.6	73.0
IRB-MT	42	61.1	56.3	59.5	56.3	76.5	70.5	77.8	72.0	68.9	64.6	78.8	74.2	78.2	72.9	38.5	36.6	76.6	71.1	79.0	73.4	76.2	70.6	76.1	70.1
TranssionMT	43	67.5	60.6	59.4	57.4	68.9	58.2	71.9	63.0	–	–	–	–	–	–	38.8	36.4	71.9	64.1	78.5	71.3	72.0	65.3	–	–
*NLLB	44	70.0	63.3	62.0	58.4	77.2	68.9	79.0	71.6	68.8	63.5	74.0	68.1	77.6	71.4	23.9	22.5	76.2	69.1	23.9	22.5	75.2	67.6	64.6	59.6
DLUT_GTCOM	45	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	78.7	72.6	–	–	77.9	71.5	–	–
Yandex	46	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	78.2	72.6	–	–	–	–	–	–
*Qwen 2.5	47	54.7	50.2	61.3	58.9	65.5	59.4	49.8	46.8	42.9	40.4	76.9	72.5	71.8	67.2	36.7	33.8	72.1	67.0	61.3	56.8	63.6	59.5	76.8	71.2
CUNI-SFT	48	–	–	–	–	78.3	70.7	–	–	–	–	–	–	–	–	–	–	–	–	77.9	71.5	76.9	70.7	–	–
*ONLINE-G	49	62.9	53.5	–	–	73.2	59.9	74.4	61.6	68.6	59.7	71.9	61.2	68.2	57.3	–	–	78.8	72.2	76.8	70.7	77.1	71.0	71.9	62.9
SH	50	–	–	–	–	–	–	–	–	–	–	80.2	75.6	–	–	–	–	–	–	–	–	–	–	–	–
CGFOKUS	51	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	76.6	70.7	–	–
Kaze-MT	52	32.1	27.5	32.0	27.6	32.5	27.9	32.5	28.1	32.3	27.7	33.3	28.7	33.5	29.0	33.1	28.5	32.8	28.2	33.1	28.4	33.1	28.5	32.4	28.2
ctpc_nlp	52	32.1	27.5	32.0	27.6	62.4	53.5	32.5	28.1	32.3	27.7	33.3	28.7	33.5	29.0	33.1	28.5	32.8	28.2	33.1	28.4	33.1	28.5	32.4	28.2
KYUoM	52	32.1	27.5	32.0	27.6	32.5	27.9	32.5	28.1	32.3	27.7	33.3	28.7	33.5	29.0	33.1	28.5	32.8	28.2	33.1	28.4	63.8	54.6	32.4	28.2
CUNI-DocTransformer	55	–	–	–	–	76.3	58.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	48.6	47.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 3: CometKiwi (higher is better) scores across language directions for original and normalised inputs (manual segmentation in both instances). System are sorted by their CometKiwi-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig	norm	orig
Shy-hunyuan-MT	1	3.4	3.8	3.9	4.2	4.7	5.4	5.7	6.4	5.2	6.1	4.0	4.5	3.7	4.1	11.8	11.6	3.1	3.8	2.2	2.6	3.9	4.6	2.3	2.8
GemTrans	2	3.5	3.9	4.0	4.3	4.8	5.5	6.1	6.9	6.9	7.3	4.0	4.5	3.8	4.3	14.7	15.0	3.4	4.0	2.1	2.7	4.1	4.7	2.5	2.9
Yolu	3	3.6	4.2	5.5	5.8	5.0	5.9	5.7	6.7	9.3	11.4	4.2	4.6	4.0	4.5	9.3	11.4	3.7	4.6	2.1	2.7	4.4	5.3	2.7	3.2
CommandA-WMT	4	4.1	4.7	4.4	4.8	4.5	5.1	6.6	7.3	7.9	8.9	4.1	4.5	3.8	4.3	10.7	10.8	4.1	4.7	3.9	4.6	3.9	4.6	2.8	3.2
Laniqo	5	–	–	–	–	4.9	5.9	5.4	6.5	–	–	4.4	4.8	4.0	4.5	–	–	3.5	4.5	–	–	4.2	4.9	2.7	3.3
*GPT-4.1	6	5.5	6.0	6.4	6.6	5.6	6.6	6.6	7.4	6.1	7.0	4.4	4.8	4.2	4.7	15.0	15.0	4.3	5.1	2.6	3.2	4.9	5.7	2.9	3.4
*Gemini 2.5 Pro	7	6.2	6.5	6.2	6.5	5.9	6.8	6.7	7.6	6.0	6.9	4.4	4.9	4.5	5.0	15.7	16.0	4.6	5.4	3.0	3.3	5.2	6.1	3.0	3.5
SalamandraTA	8	7.6	8.6	9.7	10.3	5.7	7.2	6.7	8.4	6.9	8.4	6.0	6.9	6.0	6.9	–	–	4.7	6.2	2.4	3.0	5.0	6.3	3.6	4.4
Kaze-MT	9	8.9	9.3	8.4	8.8	9.4	9.7	8.7	9.2	8.6	9.0	8.7	9.0	8.8	9.1	9.3	9.7	9.1	9.5	8.4	8.7	9.2	9.5	8.8	9.1
KYUoM	10	8.9	9.3	8.4	8.8	9.4	9.7	8.7	9.2	8.6	9.0	8.7	9.0	8.8	9.1	9.3	9.7	9.1	9.5	8.4	8.7	12.5	15.3	8.8	9.1
ctpc_nlp	11	8.9	9.3	8.4	8.8	12.7	15.2	8.7	9.2	8.6	9.0	8.7	9.0	8.8	9.1	9.3	9.7	9.1	9.5	8.4	8.7	9.2	9.5	8.8	9.1
Yandex	12	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.8	4.5	–	–	–	–	–	–
NNTSU	12	–	–	–	–	–	–	–	–	–	–	4.1	4.6	–	–	–	–	–	–	–	–	–	–	–	–
KIKIS	12	–	–	–	–	–	–	–	–	–	–	4.0	4.6	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	12	–	–	–	–	–	–	–	–	5.8	7.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
UvA-MT	16	5.0	5.6	6.5	7.1	5.9	6.6	8.2	8.9	9.4	9.9	4.7	5.2	4.3	4.8	13.0	13.0	4.2	5.0	2.6	3.1	4.9	5.5	3.2	3.5
IR-MultiagentMT	17	5.5	5.8	6.7	6.9	6.2	6.8	7.7	8.3	7.3	8.0	4.8	5.3	4.6	4.9	14.7	14.9	4.6	5.2	2.8	3.2	5.3	5.8	3.2	3.6
*Gemma 3 27B	18	5.5	5.9	6.7	6.8	6.1	6.7	7.7	8.6	8.2	8.8	4.6	5.0	4.4	5.0	15.4	15.7	4.4	5.2	2.7	3.2	5.2	6.0	3.1	3.5
Algharb	19	4.2	4.7	6.3	6.5	6.2	7.0	6.9	8.0	9.3	11.4	4.7	5.0	4.6	5.1	9.3	11.4	4.9	5.8	2.9	3.5	5.5	6.4	3.2	3.8
*ONLINE-B	20	4.0	4.9	7.2	9.8	5.9	7.3	6.9	8.3	6.1	7.2	4.3	4.9	4.4	5.0	–	–	5.6	6.7	4.0	4.9	5.3	6.6	3.0	3.9
*TowerPlus 9B	21	12.9	13.3	5.3	5.8	6.4	7.7	15.8	16.4	6.2	7.2	4.8	5.3	4.7	5.2	18.1	17.7	4.8	5.7	4.6	5.3	5.2	6.3	3.3	4.0
*TowerPlus 72B	22	7.1	7.7	5.3	5.8	6.6	7.8	13.0	13.5	6.5	7.6	4.6	5.3	4.7	5.3	18.3	18.1	4.8	5.8	3.7	4.2	5.6	6.6	3.4	4.1
*DeepSeek V3	23	5.8	6.4	5.8	6.1	6.0	7.0	7.7	8.5	7.6	8.5	4.6	5.0	4.4	5.0	15.3	15.5	4.5	5.2	2.7	3.3	5.3	6.0	3.0	3.5
*Claude4	24	5.7	6.2	5.9	6.4	6.3	7.6	7.9	9.1	7.1	8.3	4.5	5.1	4.4	4.9	15.3	15.5	4.9	6.0	2.9	3.6	5.5	6.5	3.0	3.7
*Mistral-Medium	25	6.1	6.6	6.6	7.0	6.3	7.2	9.1	10.1	9.7	10.2	4.5	5.1	4.5	5.0	–	–	–	–	–	–	5.4	6.2	2.8	3.4
*CommandA	26	5.6	6.1	6.6	7.2	6.1	7.0	11.0	11.9	12.3	12.7	4.7	5.1	4.4	4.9	15.9	15.7	5.0	5.9	3.0	3.6	5.5	6.4	3.3	3.7
*Qwen3 235B	27	7.4	7.9	7.6	8.0	6.7	7.9	10.6	11.9	11.3	12.4	4.7	5.3	4.5	4.9	15.8	16.1	4.8	5.8	3.5	4.3	5.8	6.7	2.9	3.5
SRPOL	28	4.5	5.4	–	–	5.6	7.1	7.0	8.5	–	–	4.9	5.6	–	–	–	–	5.0	6.1	–	–	5.3	6.4	3.3	4.0
*AyaExpanse 8B	29	4.8	5.3	6.9	7.4	7.0	7.7	23.9	23.9	23.2	23.2	5.0	5.4	4.7	5.2	21.6	21.2	5.9	6.5	5.5	6.1	6.0	6.7	3.7	4.0
In2x	30	–	–	–	–	–	–	–	–	–	–	4.2	4.7	–	–	–	–	–	–	–	–	–	–	–	–
IRB-MT	31	6.3	6.7	7.9	8.1	6.7	7.2	8.9	9.7	10.4	10.4	4.8	5.3	4.8	5.4	14.3	14.6	4.7	5.4	2.8	3.4	5.3	6.1	3.1	3.6
*Gemma 3 12B	32	6.5	6.9	8.1	8.3	6.6	7.5	9.3	10.0	10.9	11.3	5.2	5.7	5.1	5.6	14.1	14.1	5.0	5.6	2.9	3.4	5.3	6.1	3.4	3.9
TranssionTranslate	33	6.7	8.1	7.2	9.5	6.3	8.0	7.2	8.9	6.4	7.8	4.7	5.5	4.5	5.6	9.3	11.4	4.9	6.6	4.2	7.0	5.7	7.2	3.7	4.7
Wenyiil	34	4.5	5.5	6.7	7.1	7.0	8.2	8.0	9.5	9.3	11.4	5.1	5.7	5.1	5.9	9.3	11.4	5.5	6.8	3.1	3.8	6.0	7.1	3.4	4.1
*AyaExpanse 32B	35	5.7	6.1	8.1	8.1	6.4	7.0	20.8	20.6	20.3	20.4	4.8	5.1	4.6	5.0	19.3	19.1	5.2	5.8	3.8	4.3	5.6	6.2	3.4	3.9
*EuroLLM 22B	36	5.5	6.3	5.9	6.6	6.3	7.5	7.7	8.9	21.9	21.7	4.9	5.7	4.8	5.6	19.6	18.5	5.3	6.3	2.9	3.7	5.8	6.9	3.3	4.0
AMI	37	–	–	–	–	–	–	–	–	6.9	8.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	38	6.0	6.6	6.7	7.2	6.3	7.6	7.8	9.1	8.2	9.2	4.7	5.1	4.5	5.0	15.1	15.3	4.9	6.0	2.8	3.5	5.4	6.4	3.0	3.6
*NLLB	39	7.3	8.6	6.7	7.6	7.9	10.1	10.0	11.8	9.2	10.8	8.1	9.0	6.5	7.8	12.9	13.8	7.7	9.7	12.9	13.8	7.9	9.9	7.9	8.8
DLUT_GTCOM	40	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.8	5.7	–	–	5.9	7.3	–	–
*EuroLLM 9B	41	6.3	7.0	7.6	8.4	6.7	8.1	8.1	9.5	22.3	21.8	5.3	5.8	5.1	5.8	15.8	15.1	5.7	7.0	3.2	4.4	6.1	7.3	3.6	4.4
*ONLINE-W	42	5.1	6.4	–	–	6.6	10.0	9.1	13.5	–	–	5.3	6.2	4.5	6.3	–	–	5.1	7.0	–	–	5.5	7.0	3.7	4.4
*CommandR	43	5.4	6.2	10.1	10.5	8.6	9.8	23.3	23.2	21.6	21.9	5.4	6.0	5.5	6.0	20.1	19.6	8.5	9.8	6.2	6.7	8.7	9.7	3.8	4.5
Systran	44	–	–	–	–	–	–	–	–	–	–	4.5	5.2	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-SFT	45	–	–	–	–	7.2	9.3	–	–	–	–	–	–	–	–	–	–	–	–	3.3	4.4	6.3	7.8	–	–
CUNI-MH-v2	46	–	–	–	–	6.3	7.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama 3.1 8B	47	10.3	10.7	8.9	9.2	9.3	10.4	15.5	15.7	17.8	17.8	6.4	6.7	6.4	6.9	16.6	16.4	7.7	8.7	4.0	4.6	8.3	9.1	3.8	4.4
*Qwen 2.5	48	11.0	11.5	11.7	11.8	11.9	12.8	22.0	22.0	22.8	22.8	6.8	7.4	7.5	7.9	20.6	20.0	7.6	8.6	5.8	6.7	12.0	12.7	3.6	4.2
CGFOKUS	49	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.9	6.9	–	–
bb88	49	–	–	–	–	–	–	–	–	–	–	4.9	5.4	–	–	–	–	–	–	–	–	–	–	–	–
TranssionMT	51	6.8	8.9	7.2	8.5	12.3	16.0	14.2	17.0	–	–	–	–	–	–	15.9	16.7	10.0	12.3	3.5	5.0	9.4	11.0	–	–
*ONLINE-G	52	10.3	13.4	–	–	9.9	14.7	12.2	16.5	11.1	15.1	9.6	12.7	9.8	12.1	–	–	5.0	7.0	4.9	6.1	6.0	7.7	6.7	8.9
*Mistral 7B	53	14.9	15.3	14.1	14.1	12.9	13.5	23.6	23.7	22.6	22.6	8.5	8.8	8.2	8.5	22.1	22.2	9.5	10.5	4.6	5.3	10.0	10.5	5.4	6.0
SH	54	–	–	–	–	–	–	–	–	–	–	5.4	6.1	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-DocTransformer	55	–	–	–	–	8.2	13.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	10.7	11.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 4: MetricX scores (lower is better) across language directions for original and normalised inputs (manual segmentation in both instances). System are sorted by their MetricX-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	ma nl	man	nl	man	nl	man	nl	
Yolu	1	73.4	72.7	68.9	66.2	76.3	75.6	79.6	78.7	22.7	23.0	79.7	78.8	79.3	78.2	22.7	23.0	76.0	75.2	78.5	78.2	75.5	74.6	76.1	75.8
CommandA-WMT	2	65.6	66.2	72.8	73.6	75.1	74.8	77.8	77.8	70.0	68.9	78.5	78.2	78.1	77.8	62.6	62.0	74.7	74.4	75.0	75.1	75.3	74.7	74.7	74.5
SalamandraTA	3	66.8	65.4	57.4	56.9	76.0	76.5	78.9	79.2	72.2	72.0	78.4	77.4	77.1	76.1	–	–	75.5	75.7	78.3	78.5	75.1	75.3	75.4	74.8
Shy-hunyuan-MT	4	70.7	70.2	76.6	76.6	74.8	74.4	77.6	77.2	71.4	71.8	77.7	77.7	76.9	76.8	52.2	50.5	73.9	73.6	77.0	76.8	73.7	73.8	73.9	73.9
Laniqo	5	–	–	–	–	74.7	73.8	78.0	77.4	–	–	78.3	77.2	78.1	76.9	–	–	74.7	73.8	–	–	74.6	73.8	74.7	73.9
SRPOL	6	71.2	71.6	–	–	74.4	74.5	77.4	77.2	–	–	78.5	78.4	–	–	–	–	74.1	74.2	–	–	73.8	73.8	74.2	74.2
UvA-MT	7	66.4	66.4	68.2	68.6	73.8	73.7	75.7	75.5	67.9	67.6	78.2	78.0	77.6	77.4	37.1	37.4	73.8	73.4	76.7	76.5	73.5	73.3	74.4	74.3
*ONLINE-B	8	71.4	71.1	54.5	54.3	73.1	72.8	75.8	75.6	71.4	71.2	78.3	78.2	76.6	76.4	–	–	73.3	72.9	74.5	74.4	72.1	72.0	73.8	73.6
*TowerPlus 9B	9	46.0	45.2	75.8	75.7	72.9	73.2	57.9	59.2	70.8	70.8	77.5	77.4	77.2	77.1	41.3	42.0	73.7	73.2	60.2	60.6	72.8	72.6	73.6	73.5
TranssionTranslate	10	61.7	60.2	54.9	54.6	72.9	72.9	76.0	75.7	71.1	71.1	78.5	78.3	77.1	76.9	22.7	23.0	73.4	73.3	65.3	64.8	71.7	71.8	73.8	73.5
*TowerPlus 72B	11	60.4	60.7	76.2	76.0	72.7	72.8	67.4	67.5	70.4	70.3	77.1	77.1	77.0	76.9	38.7	42.1	73.4	73.4	68.8	69.3	72.5	72.4	74.1	73.9
GemTrans	12	70.3	69.9	76.9	76.6	72.3	72.2	74.8	74.9	67.5	67.3	76.2	75.9	75.7	75.3	34.3	34.7	72.4	72.5	75.1	74.9	72.7	72.6	73.1	72.7
*AyaExpanse 32B	13	60.5	60.3	62.8	62.5	72.7	72.2	50.9	50.9	47.8	47.8	77.0	76.9	76.4	76.2	43.2	44.0	73.1	72.6	68.3	67.6	72.4	72.4	72.6	72.7
*Llama 3.1 8B	14	52.0	51.8	62.1	62.1	66.7	66.7	62.6	62.6	53.4	53.2	72.2	71.9	72.0	71.8	48.8	48.9	69.2	69.2	68.1	68.0	68.2	68.1	70.9	70.5
Systran	15	–	–	–	–	–	–	–	–	–	–	81.0	80.6	–	–	–	–	–	–	–	–	–	–	–	–
In2x	15	–	–	–	–	–	–	–	–	–	–	78.7	78.4	–	–	–	–	–	–	–	–	–	–	–	–
*GPT-4.1	17	57.9	57.8	59.6	59.3	73.2	73.1	76.5	76.3	70.5	70.5	77.2	77.0	76.8	76.5	32.2	32.4	72.6	72.4	76.2	76.1	72.5	72.4	73.1	73.0
*CommandA	18	59.3	59.1	60.2	60.2	73.5	73.4	71.3	71.0	63.7	63.9	77.7	77.8	77.4	77.2	39.9	40.3	73.1	73.0	74.5	74.4	72.5	72.4	73.8	73.6
*EuroLLM 22B	19	66.0	65.0	73.6	73.2	72.5	72.3	76.1	76.0	43.1	42.5	76.7	76.4	75.9	75.5	36.7	37.1	73.0	72.6	74.2	74.0	72.1	72.2	73.1	72.8
*AyaExpanse 8B	20	69.7	69.5	71.1	71.2	71.4	71.4	37.5	37.4	37.4	37.4	76.7	76.2	76.2	75.8	38.2	38.3	72.0	71.9	56.8	56.7	71.9	71.7	72.5	72.3
*Claude4	21	59.4	59.3	59.5	59.4	72.8	72.9	74.9	74.8	69.4	69.3	77.5	77.4	76.9	76.6	37.6	37.7	72.3	72.2	75.3	75.2	72.2	72.2	73.0	72.9
*DeepSeek V3	22	58.9	59.3	62.5	62.3	72.4	72.4	75.3	75.5	68.8	68.7	76.4	76.2	75.7	75.7	35.4	35.6	72.2	71.6	75.4	75.4	71.7	71.3	70.5	70.2
IR-MultiagentMT	23	63.7	63.4	61.5	61.7	72.5	72.4	74.9	74.9	69.6	69.2	76.4	76.6	76.5	76.6	36.2	36.2	73.0	72.2	75.2	75.5	72.5	72.6	72.9	72.8
*CommandR	24	66.5	66.7	62.2	62.7	67.8	67.8	39.5	40.4	38.7	38.7	75.3	75.2	74.6	73.8	41.2	41.6	66.1	65.9	57.2	57.0	66.9	66.9	71.4	70.9
*Gemma 3 12B	25	57.5	57.3	57.2	57.6	71.5	71.8	72.7	72.6	64.3	64.2	75.5	75.4	74.8	74.5	42.1	42.3	72.6	72.1	74.0	73.9	72.4	71.8	72.5	72.2
Algharb	26	69.2	69.0	56.7	56.2	71.6	71.2	74.9	74.9	22.7	23.0	75.8	75.4	75.1	74.6	22.7	23.0	70.6	70.3	74.7	74.4	70.9	70.7	71.3	71.1
*ONLINE-W	27	68.7	68.3	–	–	68.9	67.3	68.0	66.2	–	–	73.8	69.6	75.4	74.4	–	–	72.5	71.4	–	–	72.1	72.0	73.0	70.8
*Gemma 3 27B	28	59.1	59.2	59.8	59.4	72.5	72.3	75.1	75.1	68.1	68.0	76.5	76.1	75.4	75.3	32.6	32.7	72.4	72.3	75.0	74.9	72.5	72.2	72.5	72.5
*EuroLLM 9B	29	64.7	64.8	65.4	66.9	72.4	72.0	75.0	74.8	41.5	41.3	75.9	75.9	75.6	75.3	28.5	29.6	72.3	71.7	70.6	70.9	71.6	71.3	72.0	71.6
Wenyiil	30	67.6	68.3	55.6	55.6	68.7	69.2	72.4	73.2	22.7	23.0	74.3	74.6	73.3	73.7	22.7	23.0	69.0	69.4	72.8	73.5	69.8	69.6	70.4	70.5
*Mistral 7B	31	45.0	44.3	52.8	52.4	59.4	59.2	39.1	39.1	40.4	40.5	67.9	67.7	67.6	67.2	40.1	40.2	66.2	66.1	64.6	63.4	65.4	65.2	65.9	65.7
CUNI-MH-v2	32	–	–	–	–	73.0	72.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
NNTSU	32	–	–	–	–	–	–	–	–	–	–	78.2	77.6	–	–	–	–	–	–	–	–	–	–	–	–
bb88	32	–	–	–	–	–	–	–	–	–	–	77.8	77.4	–	–	–	–	–	–	–	–	–	–	–	–
AMI	32	–	–	–	–	–	–	–	–	69.4	69.9	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	32	–	–	–	–	–	–	–	–	69.1	69.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	37	57.7	57.5	57.3	57.3	71.8	71.6	74.4	74.7	67.0	66.6	75.8	75.5	75.4	75.3	34.8	34.5	72.6	72.4	74.7	74.5	72.3	72.2	72.4	72.0
*Gemini 2.5 Pro	38	54.8	54.5	56.5	56.6	70.7	70.6	74.4	74.5	68.8	68.7	75.1	75.1	74.0	73.6	35.2	35.1	69.7	69.9	74.3	74.2	70.2	69.9	70.2	70.1
KIKIS	39	–	–	–	–	–	–	–	–	–	–	77.3	77.1	–	–	–	–	–	–	–	–	–	–	–	–
*Mistral-Medium	40	60.3	59.8	60.6	60.5	72.5	72.4	73.3	73.1	66.9	66.8	76.7	76.7	76.2	75.8	–	–	–	–	–	–	72.0	72.1	72.8	72.7
*Qwen3 235B	41	59.0	59.6	60.4	60.2	71.3	71.3	69.7	69.9	63.5	63.3	76.9	76.7	76.0	75.5	36.4	36.6	72.2	72.1	72.5	72.8	71.2	71.2	73.0	73.0
IRB-MT	42	56.3	56.0	56.3	56.7	70.5	69.9	72.0	72.0	64.6	63.9	74.2	74.1	72.9	72.6	36.6	37.4	71.1	70.1	73.4	73.1	70.6	70.1	70.1	69.9
TranssionMT	43	60.6	60.3	57.4	57.1	58.2	57.5	63.0	62.0	–	–	–	–	–	–	36.4	35.6	64.1	63.0	71.3	71.0	65.3	64.5	–	–
*NLLB	44	63.3	60.7	58.4	58.5	68.9	63.6	71.6	66.5	63.5	61.3	68.1	64.6	71.4	67.5	22.5	23.8	69.1	64.2	22.5	23.8	67.6	64.4	59.6	57.4
DLUT_GTCOM	45	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	72.6	72.5	–	–	71.5	71.8	–	–
Yandex	46	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	72.6	72.3	–	–	–	–	–	–
*Qwen 2.5	47	50.2	51.0	58.9	58.5	59.4	59.2	46.8	47.4	40.4	40.4	72.5	72.3	67.2	67.2	33.8	34.5	67.0	66.8	56.8	57.0	59.5	59.5	71.2	70.7
CUNI-SFT	48	–	–	–	–	70.7	70.9	–	–	–	–	–	–	–	–	–	–	–	–	71.5	72.3	70.7	70.6	–	–
*ONLINE-G	49	53.5	53.1	–	–	59.9	59.4	61.6	61.6	59.7	58.8	61.2	60.0	57.3	56.1	–	–	72.2	71.8	70.7	70.6	71.0	70.6	62.9	61.6
SH	50	–	–	–	–	–	–	–	–	–	–	75.6	74.8	–	–	–	–	–	–	–	–	–	–	–	–
CGFOKUS	51	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	70.7	70.8	–	–
Kaze-MT	52	27.5	25.0	27.6	25.0	27.9	25.3	28.1	25.5	27.7	25.2	28.7	26.4	29.0	26.6	28.5	26.0	28.2	25.7	28.4	26.0	28.5	26.0	28.2	25.8
ctpc_nlp	52	27.5	25.0	27.6	25.0	53.5	55.1	28.1	25.5	27.7	25.2	28.7	26.4	29.0	26.6	28.5	26.0	28.2	25.7	28.4	26.0	28.5	26.0	28.2	25.8
KYUoM	52	27.5	25.0	27.6	25.0	27.9	25.3	28.1	25.5	27.7	25.2	28.7	26.4	29.0	26.6	28.5	26.0	28.2	25.7	28.4	26.0	54.6	55.3	28.2	25.8
CUNI-DocTransformer	55	–	–	–	–	58.6	57.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	47.0	47.8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 5: CometKiwi scores (higher is better) across languages directions for manual (man) segmentation versus newline (nl) segmentation on original outputs. System are sorted by their CometKiwi-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

System	Rank	en-ar_EG		en-bho_IN		en-cs_CZ		en-et_EE		en-is_IS		en-ja_JP		en-ko_KR		en-mas_KE		en-ru_RU		en-sr_Latn_RS		en-uk_UA		en-zh_CN	
		man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl	man	nl
Shy-hunyuan-MT	1	3.8	4.0	4.2	4.3	5.4	5.5	6.4	6.4	6.1	6.0	4.5	4.6	4.1	4.2	11.6	11.8	3.8	4.0	2.6	2.7	4.6	4.6	2.8	2.9
GemTrans	2	3.9	4.0	4.3	4.4	5.5	5.5	6.9	6.8	7.3	7.3	4.5	4.6	4.3	4.3	15.0	14.8	4.0	4.1	2.7	2.8	4.7	4.8	2.9	3.1
Yolu	3	4.2	4.2	5.8	6.2	5.9	5.9	6.7	6.8	11.4	11.4	4.6	4.7	4.5	4.5	11.4	11.4	4.6	4.5	2.7	2.8	5.3	5.2	3.2	3.1
CommandA-WMT	4	4.7	4.6	4.8	4.8	5.1	5.2	7.3	7.2	8.9	8.9	4.5	4.6	4.3	4.4	10.8	10.9	4.7	4.8	4.6	4.6	4.6	4.6	3.2	3.3
Laniqo	5	–	–	–	–	5.9	5.7	6.5	6.5	–	–	4.8	5.0	4.5	4.6	–	–	4.5	4.4	–	–	4.9	4.9	3.3	3.2
*GPT-4.1	6	6.0	6.1	6.6	6.8	6.6	6.6	7.4	7.4	7.0	7.0	4.8	4.9	4.7	4.8	15.0	14.8	5.1	5.2	3.2	3.2	5.7	5.7	3.4	3.5
*Gemini 2.5 Pro	7	6.5	6.5	6.5	6.5	6.8	6.8	7.6	7.6	6.9	6.9	4.9	4.9	5.0	5.1	16.0	15.8	5.4	5.5	3.3	3.5	6.1	6.1	3.5	3.5
SalamandraTA	8	8.6	9.0	10.3	10.4	7.2	6.9	8.4	8.1	8.4	8.3	6.9	7.1	6.9	7.1	–	–	6.2	6.1	3.0	3.0	6.3	6.1	4.4	4.5
Kaze-MT	9	9.3	10.6	8.8	10.0	9.7	11.0	9.2	10.5	9.0	10.2	9.0	10.2	9.1	10.4	9.7	11.0	9.5	10.7	8.7	10.1	9.5	10.8	9.1	10.3
KYUoM	10	9.3	10.6	8.8	10.0	9.7	11.0	9.2	10.5	9.0	10.2	9.0	10.2	9.1	10.4	9.7	11.0	9.5	10.7	8.7	10.1	15.3	15.4	9.1	10.3
ctpc_nlp	11	9.3	10.6	8.8	10.0	15.2	14.2	9.2	10.5	9.0	10.2	9.0	10.2	9.1	10.4	9.7	11.0	9.5	10.7	8.7	10.1	9.5	10.8	9.1	10.3
Yandex	12	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.5	4.6	–	–	–	–	–	–
NNTSU	12	–	–	–	–	–	–	–	–	–	–	4.6	4.6	–	–	–	–	–	–	–	–	–	–	–	–
KIKIS	12	–	–	–	–	–	–	–	–	–	–	4.6	4.7	–	–	–	–	–	–	–	–	–	–	–	–
Erlendur	12	–	–	–	–	–	–	–	–	7.1	7.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–
UvA-MT	16	5.6	5.5	7.1	7.1	6.6	6.6	8.9	8.8	9.9	10.0	5.2	5.2	4.8	4.9	13.0	13.0	5.0	5.0	3.1	3.1	5.5	5.6	3.5	3.6
IR-MultiagentMT	17	5.8	6.0	6.9	7.0	6.8	6.9	8.3	8.4	8.0	8.2	5.3	5.3	4.9	5.0	14.9	14.9	5.2	5.4	3.2	3.3	5.8	5.8	3.6	3.7
*Gemma 3 27B	18	5.9	5.9	6.8	7.0	6.7	6.8	8.6	8.5	8.8	8.9	5.0	5.1	5.0	5.1	15.7	15.8	5.2	5.3	3.2	3.3	6.0	6.1	3.5	3.6
Algharb	19	4.7	4.6	6.5	6.6	7.0	6.8	8.0	7.6	11.4	11.4	5.0	5.0	5.1	5.0	11.4	11.4	5.8	5.5	3.5	3.5	6.4	6.2	3.8	3.7
*ONLINE-B	20	4.9	5.0	9.8	9.8	7.3	7.4	8.3	8.4	7.2	7.2	4.9	5.0	5.0	5.1	–	–	6.7	6.8	4.9	5.0	6.6	6.7	3.9	4.1
*TowerPlus 9B	21	13.3	13.4	5.8	5.9	7.7	7.6	16.4	16.0	7.2	7.2	5.3	5.3	5.2	5.3	17.7	17.8	5.7	5.9	5.3	5.4	6.3	6.3	4.0	4.1
*TowerPlus 72B	22	7.7	7.8	5.8	5.8	7.8	7.7	13.5	13.3	7.6	7.5	5.3	5.3	5.3	5.3	18.1	17.9	5.8	5.9	4.2	4.2	6.6	6.6	4.1	4.1
*DeepSeek V3	23	6.4	6.3	6.1	6.2	7.0	6.9	8.5	8.5	8.5	8.5	5.0	5.1	5.0	5.1	15.5	15.4	5.2	5.4	3.3	3.3	6.0	6.0	3.5	3.6
*Claude4	24	6.2	6.2	6.4	6.4	7.6	7.4	9.1	9.0	8.3	8.3	5.1	5.2	4.9	5.0	15.5	15.4	6.0	6.0	3.6	3.6	6.5	6.5	3.7	3.7
*Mistral-Medium	25	6.6	6.7	7.0	6.9	7.2	7.1	10.1	10.1	10.2	10.2	5.1	5.1	5.0	5.0	–	–	–	–	–	–	6.2	6.2	3.4	3.4
*CommandA	26	6.1	6.2	7.2	7.1	7.0	6.9	11.9	11.8	12.7	12.5	5.1	5.1	4.9	4.9	15.7	15.7	5.9	5.9	3.6	3.7	6.4	6.4	3.7	3.8
*Qwen3 235B	27	7.9	7.9	8.0	8.1	7.9	7.7	11.9	11.7	12.4	12.1	5.3	5.3	4.9	5.0	16.1	16.1	5.8	5.9	4.3	4.2	6.7	6.7	3.5	3.5
SRPOL	28	5.4	5.3	–	–	7.1	6.9	8.5	8.2	–	–	5.6	5.5	–	–	–	–	6.1	6.0	–	–	6.4	6.2	4.0	4.0
*AyaExpanse 8B	29	5.3	5.3	7.4	7.4	7.7	7.8	23.9	23.7	23.2	23.2	5.4	5.6	5.2	5.2	21.2	20.9	6.5	6.5	6.1	6.1	6.7	6.7	4.0	4.2
In2x	30	–	–	–	–	–	–	–	–	–	–	4.7	4.7	–	–	–	–	–	–	–	–	–	–	–	–
IRB-MT	31	6.7	6.7	8.1	8.0	7.2	7.3	9.7	9.6	10.4	10.6	5.3	5.4	5.4	5.5	14.6	14.5	5.4	5.4	3.4	3.5	6.1	6.1	3.6	3.7
*Gemma 3 12B	32	6.9	7.0	8.3	8.3	7.5	7.4	10.0	10.0	11.3	11.3	5.7	5.6	5.6	5.7	14.1	14.1	5.6	5.7	3.4	3.6	6.1	6.3	3.9	4.0
TranssionTranslate	33	8.1	8.1	9.5	9.7	8.0	7.6	8.9	8.6	7.8	7.5	5.5	5.5	5.6	5.5	11.4	11.4	6.6	6.4	7.0	7.0	7.2	6.9	4.7	4.7
Wenyiil	34	5.5	5.0	7.1	7.1	8.2	7.7	9.5	8.6	11.4	11.4	5.7	5.4	5.9	5.5	11.4	11.4	6.8	6.0	3.8	3.6	7.1	6.7	4.1	3.8
*AyaExpanse 32B	35	6.1	6.1	8.1	8.1	7.0	7.1	20.6	20.9	20.4	20.5	5.1	5.2	5.0	5.0	19.1	19.0	5.8	5.8	4.3	4.3	6.2	6.2	3.9	3.9
*EuroLLM 22B	36	6.3	6.4	6.6	6.6	7.5	7.3	8.9	8.8	21.7	21.7	5.7	5.8	5.6	5.7	18.5	18.2	6.3	6.4	3.7	3.7	6.9	6.8	4.0	4.1
AMI	37	–	–	–	–	–	–	–	–	8.2	8.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama-4-Maverick	38	6.6	6.4	7.2	7.1	7.6	7.5	9.1	8.8	9.2	9.0	5.1	5.2	5.0	5.0	15.3	15.4	6.0	5.9	3.5	3.5	6.4	6.4	3.6	3.7
*NLLB	39	8.6	9.2	7.6	7.7	10.1	11.2	11.8	13.3	10.8	13.1	9.0	9.7	7.8	8.6	13.8	12.0	9.7	11.6	13.8	12.0	9.9	11.0	8.8	9.2
DLUT_GTCOM	40	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.7	5.2	–	–	7.3	7.0	–	–
*EuroLLM 9B	41	7.0	7.1	8.4	8.0	8.1	8.1	9.5	9.5	21.8	21.7	5.8	5.9	5.8	5.9	15.1	14.2	7.0	7.1	4.4	4.4	7.3	7.3	4.4	4.4
*ONLINE-W	42	6.4	6.5	–	–	10.0	10.8	13.5	14.1	–	–	6.2	6.8	6.3	6.5	–	–	7.0	7.4	–	–	7.0	7.0	4.4	4.6
*CommandR	43	6.2	6.2	10.5	10.1	9.8	9.7	23.2	22.8	21.9	21.3	6.0	6.0	6.0	6.1	19.6	18.4	9.8	9.7	6.7	6.7	9.7	9.5	4.5	4.5
Systran	44	–	–	–	–	–	–	–	–	–	–	5.2	5.2	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-SFT	45	–	–	–	–	9.3	8.7	–	–	–	–	–	–	–	–	–	–	–	–	4.4	4.3	7.8	7.5	–	–
CUNI-MH-v2	46	–	–	–	–	7.8	7.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
*Llama 3.1 8B	47	10.7	10.6	9.2	9.0	10.4	10.3	15.7	15.6	17.8	17.6	6.7	6.6	6.9	7.0	16.4	15.8	8.7	8.6	4.6	4.5	9.1	8.8	4.4	4.4
*Qwen 2.5	48	11.5	11.2	11.8	11.9	12.8	12.8	22.0	21.7	22.8	22.6	7.4	7.5	7.9	7.8	20.0	19.5	8.6	8.4	6.7	6.6	12.7	12.5	4.2	4.3
CGFOKUS	49	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6.9	6.8	–	–
bb88	49	–	–	–	–	–	–	–	–	–	–	5.4	5.2	–	–	–	–	–	–	–	–	–	–	–	–
TranssionMT	51	8.9	8.9	8.5	8.7	16.0	16.1	17.0	17.0	–	–	–	–	–	–	16.7	16.4	12.3	12.4	5.0	5.2	11.0	11.1	–	–
*ONLINE-G	52	13.4	13.5	–	–	14.7	14.8	16.5	16.5	15.1	15.5	12.7	12.9	12.1	12.2	–	–	7.0	7.1	6.1	6.1	7.7	7.8	8.9	9.2
*Mistral 7B	53	15.3	15.1	14.1	14.4	13.5	13.2	23.7	23.7	22.6	22.5	8.8	8.7	8.5	8.5	22.2	22.2	10.5	10.2	5.3	5.4	10.5	10.3	6.0	6.0
SH	54	–	–	–	–	–	–	–	–	–	–	6.1	6.1	–	–	–	–	–	–	–	–	–	–	–	–
CUNI-DocTransformer	55	–	–	–	–	13.6	14.2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
COILD-BHO	55	–	–	11.0	11.0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 6: MetricX scores (lower is better) across languages directions for manual (man) segmentation versus newline (nl) segmentation on original inputs. System are sorted by their MetricX-based ranking (over all language pairs). Systems run by the task organisers are marked with an asterisk.

Automated Evaluation for Terminology Translation Related to the EEA Agreement

Selma Dís Hauksdóttir
The University of Iceland
Reykjavík, Iceland
sdh22@hi.is

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
steinthor.steingrimsson@arnastofnun.is

Abstract

This paper presents a submission to the WMT25 test suite subtask, focusing on terminology in official documents related to the EEA Agreement. The test suite evaluates the accuracy of MT systems in translating terminology for the English→Icelandic translation direction when applied to EEA documents. We focus on the use of terminology in four domains of the agreement; science, technology, economics, and society. We find that manual evaluation confirms that our test suite can be helpful in selecting the best MT system for working with these domains. Surprisingly, an online system which does not achieve very high scores on the general translation task, according to preliminary results, is most adept at translating the terminology our test suite evaluates. The test suite and evaluation code are openly available on Github: https://github.com/steinst/WMT25_EEA_test_suite

1 Introduction

Pozzo (2020) argues that the law of the European Community has a multilingual framework, with 24 official languages, which calls for the use of descriptive language to maintain equal distance between each language, with legal terminology being culture-bound. The EEA Agreement is translated into two additional languages; Icelandic and Norwegian. The meaning associated with legal texts is often disputed, even in monolingual texts. The target text should perform at the same function as the source text, since they're equivalent in legal sense (Bago et al., 2022). Term inconsistencies are therefore unacceptable. Until recently, most MT systems translated documents sentence by sentence, which could result in inconsistencies in translation of terminology (see e.g. Semenov and Bojar (2022)). The Translation Centre at the Ministry for Foreign Affairs in Iceland translates around 9000 pages related to the EEA Agreement each year (Steindórsdóttir, 2022). They have been testing an MT system trained on their own corpus but the results show that the system outputs are mediocre at best. Bago et al. (2022) state that one of the main complaints of the translators is the lack of correct terminology and consistency in the MT output.

Current LLM-based systems are capable of considering larger context. In this paper we present a test suite which can help us understand if systems based on this new MT paradigm can translate the terminology correctly. We focus on terminology translations from English to Icelandic in the EEA Agreement in our submission for the WMT25 Test Suite subtask (Kocmi et al., 2025a). We evaluate 34 systems, both automatically and manually, and find that surprisingly, two out of the three highest scoring systems are Online-systems. We release our test suite and evaluation code for others to build on and to allow for further comparison between future models in this domain.

2 Related Work

Semenov and Bojar (2022) measured consistency, unambiguity and adequacy in automatically translated legal texts by counting the correct occurrences of the exact term. Gašpar et al. (2022) found that the Herfindahl-Hirshman Index (HHI) can be used to measure the consistency of terminology in a translated corpus, since the HHI works on a small amount of data.

$$\mathrm{HHI} = \sum_{i=1}^{n} S_i^2 \tag{1}$$

i ranges over n different translations for the specific term translated in the relevant text. S_i is the ratio of the number of times when the term was translated as i to the total number of times it was translated. Therefore the HHI score will be 5.0 if a term has two different translations.

$$\frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \left(\frac{f_i}{k_j} \cdot 100 \right)^2}{p} \tag{2}$$

Proceedings of the Tenth Conference on Machine Translation, pages 850–855
November 8-9, 2025

An overall translation consistency index (C_t) for a source term is calculated as follows: p is the number of translation having the source term t, and each frequency share is calculated as the ratio of its frequency f_i to the total translation occurrence within a product k_j. The score ranges between 0 and 10, with 10 being perfect consistency.

Alam et al. (2021) introduced a benchmark for evaluation of quality and consistency of terminology translation in a shared task at WMT21, with creating new evaluation datasets that were annotated by professional translators for their terminology consistency. They found that general translation quality does not have to be sacrificed for terminology compliance. Semenov et al. (2023) evaluated the efficiency of using segment-level terminology dictionaries in a shared task at WMT23, and concluded that an improvement in MT performance when using a terminology dictionary ranged between 0 and 10 ChrF points.

Two participants in the WMT 2024 Test Suite subtask (Kocmi et al., 2024) focused specifically on English→Icelandic translations: Friðriksdóttir (2024) focuses on various aspects of gender-inclusive translation, including LGBTQIA+ terminology and whether translations are current and culturally appropriate, as the terminology in that domain has been updated repeatedly. Ármannsson et al. (2024) focus on idiomatic expressions and proper names in their test suite. Their evaluation is keyword-based, checking if content words in the idioms are translated correctly and whether the proper names have correct translations.

3 Methodology

We built an automated keyword-based evaluation regarding the EEA Agreement. The aim was to test the ability of MT systems when it comes to the translation of terms in the Agreement, as disambiguation is important when it comes to the translation of legal texts. To test this, we ran the automated evaluation and confirmed our method with manual evaluation. We manually collect sentences from EU regulations and directives relevant to the EEA agreement. For each sentence, we tag terms and find the standard Icelandic translation for each term on the official Icelandic website for the EEA Agreement.[1] The Icelandic translations are used for automatic evaluation and a subset of the translations from each MT system is manually evaluated in order to understand whether the automatic evaluation is close to human judgment.

3.1 Test Suite Compilation

The terms were manually extracted from 32 EU Regulations and two EU Directives that were translated into Icelandic and published in 2024 and 2025. The aim was not to test the regulations, but to collect a diverse and descriptive sample of keywords that appear in the EEA Agreement. In some cases, we added a simple verb phrase if necessary, to build a coherent sentence. The terms were divided into four subgroups: science, technology, economics, and society, with as little overlay as possible. The subgroups are based on the groups at the Translation Centre, where the EEA Agreement is divided into said subgroups. Every sentence contained at least one term that we tested, but we did not test every term in each sentence, especially not recurring terms, since we were not testing consistency in this test suite. We gathered every word form of the terms and used it for the automatic evaluation.

The sentences were exported as a txt-file, which was sent to the test suite subtask of WMT25. Once we got the translated sentences from the 34 MT systems, we ran the automatic evaluation, see 3.2. We manually evaluated translation of the terms in 50 sentences for each system to test our automatic evaluation method, see 3.3.

Due to an error in the layout of the txt-file, some of our sentences were split, so we ended up with 256 sentences and 408 keywords. We disregard the erroneous sentences in the input and report only on the error-free ones. We have however published a corrected version on Github along with the test suite and evaluation codes, for others to build on and compare other models.

3.2 Automatic Evaluation

The automatic evaluation is keyword-based. For each MT system we check all 256 complete sentences and disregard the ones that were split up before submission. The check inspects if a given translation contains the Icelandic terms, by comparing the translation to all possible inflectional forms of the term. We look up the word forms in DIM, the Database of Icelandic Morphology (Bjarnadóttir et al., 2019), and if they are not found there, we manually create a list of acceptable forms. If the translation contains the term in any form accepted in our lists we count that as correct.

[1] https://gagnagrunnur.ees.is/

3.3 Manual Evaluation

We manually evaluated 1700 translations, 50 for each system. The sentences were chosen randomly from the complete set of 256 translations and for all systems we evaluated translations of the same 50 sentences. The evaluator is a PhD student in Translation Studies and a former translator at the Translation Centre, with a three year background as a professional translator. The manual evaluation took around 10 hours, since the focus was only on the keywords and not the sentences. One point was given for every term that was correctly translated, and the maximum points available correlated therefore with the number of terms. A point was given for acceptable translations, other than the ones that were included in our keyword list. 117 other translations were accepted, mainly synonyms, and rephrasing of terms consisting of more than one word. Additionally, we inspect the ratio of sentences that have all terms correctly translated.

4 Results

Results of the automatic evaluation are presented in Table 1 and manual evaluation in Table 2. While the main difference between the evaluation approaches is that the manual evaluation paints a picture where many of the MT systems seem to be quite adept at dealing with EEA terminology, achieving up to almost 80% accuracy, the accuracy being the ratio of terms correctly translated according to the human annotator. The automatic evaluation gives substantially lower scores, which may indicate that the keyword and word inflection lists used for the automatic evaluation are sometimes lacking. Even though that is the case, the order of the systems is very similar in the two evaluations, manual and automatic. Our main takeaway from comparison is thus that if we trust our manual evaluation to be reliable and can use that to help us select the best MT system to help us with EEA translations, we can also trust the automatic evaluation using our test suite, as the order of the system is almost identical, with the same systems being in the top 3 seats of both lists. If we compare our results to the preliminary rankings for the WMT25 general translation task (Kocmi et al., 2025b), we find that our order of systems is quite different. The most surprising results are that the system achieving first place on both our lists, ONLINE-G, is actually a low scoring system in the general translation task, ending up in 24th place out of 33 systems. We wonder why this is and speculate whether this might be an encoder-decoder system that actually contains EEA texts in their training data. If an evaluated system is trained on data from the domain being evaluated, possibly containing the same or very similar structures as being evaluated in the test suite, this data leakage can lead to overestimation of the models capabilities, see e.g. Zhu et al. (2024) and Zeng et al. (2024). This could explain why the system is particularly good in this task, but not in the general translation task where LLMs seem to have an advantage. This is not necessarily a far-fetched idea, as a substantial part of ParIce (Barkarson and Steingrímsson, 2019; Steingrímsson and Barkarson, 2021), a parallel English-Icelandic parallel corpus, comprises data from EEA-documents and this corpus is among those distributed on OPUS[2]. Other top scoring systems, on the other hand, are all in the top seats in the preliminary system ranking table.

5 Conclusions and Future Work

We evaluated 34 MT systems using our test suite, automatically and manually. The evaluation shows that while a few of the systems translate the majority of the terms correctly, they are all quite far from perfect. Our automatic evaluation orders the systems in a similar way to the manual evaluation, indicating that an automatic approach such as this one can be useful to help translators find the most useful system for the task.

A larger set of sentences and terminology would improve our test suite, especially if we include terminology from other subdomains. Given the large amount of published documents relating to the EEA Agreement it is almost, if not entirely, impossible to test for every single term that appears in those documents. We plan to look into the frequency of terms in order to reconstruct the test suite in a way that may be more indicative of real-world usage, on the one hand giving terms that appear often in the Agreement more weight, but on the other make sure that a representative part of terms that rarely appear is also included. To build further on the test suite, we also plan to look into results for each subdomain and see if MT systems perform better for a specific domain. Another interesting area is to test the translation of neologism, as acts about new developments, especially scientific and technological ones, often call for new terminology. Finally, looking into the translation of terms that have more

[2] https://opus.nlpl.eu/

System	Term Acc. (%)	Sentence Acc. (%)
ONLINE-G	55.9	42.2
Erlendur	53.3	41.0
Gemini-2.5-Pro	46.5	30.5
ONLINE-B	46.5	33.2
TranssionTranslate	46.2	32.8
hybrid	38.7	26.6
Claude-4	38.5	27.7
SalamandraTA	38.5	25.0
TowerPlus-9B	37.5	23.8
Shy	36.8	22.7
GPT-4.1	34.9	23.8
TowerPlus-72B	32.7	20.7
DeepSeek-V3	30.0	17.2
AMI	27.8	17.2
Llama-4-Maverick	27.8	17.6
NLLB	26.6	14.8
CommandA-MT	24.9	14.8
IR-MultiagentMT	24.2	14.5
Gemma-3-27B	20.1	8.6
Mistral-Medium	18.9	9.4
GemTrans	16.5	8.2
IRB-MT	13.8	6.3
UvA-MT	13.1	5.9
Gemma-3-12B	11.6	4.3
Qwen3-235B	10.9	3.9
CommandA	7.5	2.0
Llama-3.1-8B	3.1	0.8
AyaExpanse-32B	2.7	0.8
Qwen2.5-7B	0.7	0.0
CommandR7B	0.5	0.0
EuroLLM-9B	0.5	0.4
EuroLLM-22B	0.2	0.0
Mistral-7B	0.2	0.0
AyaExpanse-8B	0.0	0.0

Table 1: Automatic evaluation of the systems.

System	Term Acc. (%)	Sentence Acc. (%)
ONLINE-G	79.6	64
Erlendur	76.3	64
Gemini-2.5-Pro	72	62
TranssionTranslate	71	60
ONLINE-B	67.7	54
Claude-4	60.2	46
Shy	60.2	44
hybrid	59.1	48
TowerPlus-9B	59.1	44
GPT-4.1	57	44
SalamandraTA	55.9	42
IR-MultiagentMT	44.1	30
TowerPlus-72B	44.1	26
NLLB	43	26
DeepSeek-V3	40.9	20
AMI	37.6	26
CommandA-MT	37.6	22
Llama-4-Maverick	35.5	18
Gemma-3-27B	31.2	14
Mistral-Medium	30.1	10
GemTrans	26.9	16
Gemma-3-12B	25.8	16
UvA-MT	25.8	14
IRB-MT	24.7	14
Qwen3-235B	16.1	4
CommandA	12.9	2
Llama-3.1-8B	5.4	4
AyaExpanse-32B	2.2	0
AyaExpanse-8B	1.1	0
CommandR7B	1.1	0
EuroLLM-22B	1.1	0
EuroLLM-9B	0	0
Mistral-7B	0	0
Qwen2.5-7B	0	0

Table 2: Manual evaluation of the systems.

than one allowed Icelandic translation, based on context and subgroups, could help us understand problems that translators might miss and special attention has to be paid to.

This test suite can be adapted to other languages with relative ease, which allows further work on other language directions.

6 Limitations

This work did not consider consistency especially, which would be a logical next step, to check whether the terms are consistently translated in the same way, or whether for some MT systems correct translations may be fortuitous incidents.

Our selection of terms was not always systematic and we do not always consider all terms in a given sentence. We were not able to check every category under each subgroup in this test suite due to the size limitations.

Due to time limits we were not able to add the accepted translations, and all the word forms of said translations, from the manual evaluation to the list of accepted terms. The keyword and word inflection lists are therefore lacking for the automatic

evaluation.

As mentioned above, some sentences were split up and we ended therefore with fewer sentences than anticipated, and therefore a smaller test suite. Both the submitted test suite and a fixed one are available in the Github repository.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT Shared Task on Machine Translation Using Terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinþór Steingrímsson. 2024. Killing Two Flies with One Stone: An Attempt to Break LLMs Using English-Icelandic Idioms and Proper Names. In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458, Miami, Florida, USA. Association for Computational Linguistics.

Petra Bago, Sheila Castilho, Edoardo Celeste, Jane Dunne, Federico Gaspari, Níels Rúnar Gíslason, Andre Kåsen, Filip Klubička, Gauti Kristmannsson, Helen McHugh, and et al. 2022. Sharing high-quality language resources in the legal domain to develop neural machine translation for under-resourced European languages. *Revista de Llengua i Dret*, (78):9–34.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 140–145, Turku, Finland. Linköping University Electronic Press.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154, Turku, Finland. Linköping University Electronic Press.

Steinunn Rut Friðriksdóttir. 2024. The GenderQueer Test Suite. In *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340, Miami, Florida, USA. Association for Computational Linguistics.

Angelina Gašpar, Sanja Seljan, and Vlasta Kučiš. 2022. Measuring Terminology Consistency in Translated Corpora: Implementation of the Herfindahl-Hirshman Index. *Information*, 13(2).

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Barbara Pozzo. 2020. Looking for a Consistent Terminology in European Contract Law. *Lingue Culture Mediazioni - Languages Cultures Mediation*, 7.1:103–126.

Kirill Semenov and Ondřej Bojar. 2022. Automated Evaluation Metric for Terminology Consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Hanna Kristín Steindórsdóttir. 2022. Þýðingamiðstöð utanríkisráðuneytisins. ESB-textar og sérstaða þeirra í þýðingum.

Steinþór Steingrímsson and Starkaður Barkarson. 2021. ParIce: English-icelandic parallel corpus (21.10). CLARIN-IS.

Xianfeng Zeng, Yijin Liu, Fandong Meng, and Jie Zhou. 2024. Towards multiple references era – addressing data leakage and limited reference diversity in machine translation evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11939–11951, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Up to Par? MT Systems Take a Shot at Sports Terminology

Einar Freyr Sigurðsson, Magnús Már Magnússon, Atli Jasonarson, Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavík, Iceland
einar.freyr.sigurdsson,magnus.mar.magnusson,atli.jasonarson,
steinthor.steingrimsson@arnastofnun.is

Abstract

We present a submission to the WMT25 test suite subtask, focusing on the capabilities of MT systems to translate sports-related language. Although many sports attract extensive media attention and feature a rich, polysemous language, often shaped by active neologism and community-driven translations, the sports domain has received relatively little focus in MT research. In English-Icelandic automatic translations, sports-specific vocabulary often appears to be mistranslated. Our test suite is designed to test whether this observation holds merit. We evaluate 34 systems, both automatically and manually, and find that sports language poses challenges to a varying degree for all the systems.

1 Introduction

With the advent of large language models (LLMs), significant advances have been made in machine translation (MT) (Kocmi et al., 2024). While general translation capabilities are impressive for many high-resource and even some less-resourced languages, many MT systems commonly fail when dealing with specialized vocabulary or other rare peculiarities. We have noticed that some commonly used systems seem more prone to making errors when the topic is sports than for other common topics in the media. To investigate this further we built a test suite and submitted it to the WMT 2025 Test Suite subtask (Kocmi et al., 2025a). We compile a list of segments discussing five different sports: Basketball, football, golf, gymnastics and chess. These are international sports that are popular and have been played in Iceland for decades. As a result, these sports often have a well-established and diverse vocabulary that many speakers need to agree upon. It is therefore important that different translation systems apply vocabulary that is known and in actual use by speakers in order to translate sports texts successfully. Additionally, while "new" sports known from abroad start being played in Iceland, the vocabulary may consist of a somewhat high proportion of loanwords adapted to Icelandic (using, e.g., English-oriented stems with Icelandic inflection). Vocabulary for sports that have a long history in Iceland, however, strikes a balance between using loanwords and new (and old) words applying other methods. Our test suite is made available in plain text format with term annotations as well as evaluation code on GitHub.[1]

2 Related Work

There is a long tradition of terminology work in Iceland (see, e.g., Christensen et al. 2025). The Icelandic terminology bank (Íðorðabankinn)[2] at the Árni Magnússon Institute for Icelandic Studies is the center of terminology work in Iceland. As of August 2025, it hosts more than 70 terminologies and glossaries. Most of these give translations of the terms in one or more languages, and English is usually one of them. However, only one of said terminologies and glossaries deal specifically with sports — that is the terminology on climbing.

Furthermore, even if a given translation system may contain, e.g., material from published dictionaries in its training data, its usefulness may be limited by the fact that some well-known words from sports may be missing. Also, various words generally known by speakers of English are used in a very specific sense in sports, that may differ from the general use. Such terms may have a different translation in Icelandic altogether. An example is the noun *paint*, as in *The paint is wet*. In basketball, this refers to a specific place on a basketball court. Whereas *paint* in a general sense could be translated as 'málning' in Icelandic, in the basketball sense it can be translated as 'teigur, vítateigur', meaning something like 'free-throw lane', or even

[1] github.com/steinst/WMT25_Sports_Test_Suite
[2] idord.arnastofnun.is

Proceedings of the Tenth Conference on Machine Translation, pages 856–865
November 8-9, 2025

Sport	Segments	Words/Segment	Unique Terms	Repeated Terms
Football	100	26.7	196	94
Basketball	100	31.8	154	89
Chess	50	18.7	67	4
Gymnastics	25	20.5	46	10
Golf	25	34.5	48	14

Table 1: Number of segments, average length of segments, number of terms and how many terms are repeated in the segments. 142 unique terms are repeated in the test suite, some more than once.

as 'undir körfunni', meaning 'below/under the basket'.

We therefore find it interesting — and challenging — to look into vocabulary that may not be very well documented in lexicographic works. However, at least some of the vocabulary can be found in texts online, e.g., in corpora such as Tímarit.is[3] (e.g., Hrafnkelsson and Sævarsson 2014) and the Icelandic Gigaword Corpus[4] (Steingrímsson et al., 2018; Barkarson et al., 2022).

Knowles et al. (2023) study the performance of NMT systems at handling terminology and find that the systems are twice as likely as humans to commit terminology errors in the Hansard corpus. They measure the accuracy by counting the occurrences and translations of unique terms. In the shared task on machine translation with terminologies at WMT 2023, three evaluation approaches were used (Semenov et al., 2023): Standard metrics to evaluate general translation quality, COMET (Rei et al., 2020) and ChrF (Popović, 2015); Term success rate, or accuracy, was calculated by comparing machine-translated terms with their dictionary equivalents; Term consistency was measured to investigate whether technical terms were translated uniformly across the entire corpus.

At WMT 2024, two submissions specifically looked at problems in translating between English and Icelandic. The GenderQueer test suite (Friðriksdóttir, 2024) is built to investigate gender-inclusive translation and whether such translations are appropriate. Ármannsson et al. (2024) uses a keyword-based evaluation to check how adept MT systems are at translating idiomatic expressions and proper names.

[3] timarit.is
[4] igc.arnastofnun.is

English Segment	Icelandic Terms
She remembers when she started learning a **kip** during her first week of **bars**.	[kippur, langkippur] tvíslá

Table 2: An English segment containing sport terms, and their Icelandic counterparts.

3 Test Suite and Translations

The test suite consists of 300 segments in total, divided into five documents. The segments can be sentences or short paragraphs containing a few sentences. An example of a segment from gymnastics is shown in Table 2.

3.1 The Dataset

We decided to select popular and well-established sports for the test suite. Football (soccer in the USA) has a more than 100-year history in Iceland with the first rule book being published in 1907 (*Knattspyrnulög* 1907, Jasonarson 2025). Football is hugely popular in Iceland with a lot of coverage in the news, online and in various podcasts. Football games are broadcast live with Icelandic commentary, including games in the Icelandic divisions, games played in the biggest leagues in the world and games played by national teams. Written live commentary is also available, especially from the Icelandic leagues. There is definitely no shortage of texts if one wants to study the grammar and vocabulary of football.

We also include 100 segments of texts on basketball — another very popular sport with a lot of media coverage — whereas we have fewer segments for chess (50), golf (25) and gymnastics (25). This could have been done differently but instead of focusing on, e.g., two sports, we went for covering five, with three of them having fewer segments.

By having more text from football and basketball, we can evaluate consistency in recurring terms but there are fewer terms that occur more than once in the chess, golf and gymnastics set. Table 1 shows the number of segments, length of segments and number of terms covered for each sport represented in the test suite.

We selected sentences containing sports terminology from news reports and other sports-related coverage to ensure a varied distribution of vocabulary. Certain terms were intentionally repeated to allow for examination of model consistency. To preserve the original meaning when extracting sentences from their context, minimal edits were applied, such as shortening or rephrasing.

We mark the terms in each segment and register the Icelandic equivalent. This information is used for automatic evaluation of MT systems when applied to the test suite. Following Ármannsson et al. (2024), we carry out a keyword-based evaluation and compare the results to a manually evaluated sample to inspect whether the automatic approach judges the systems in a way close to what humans would do.

3.2 The Work Process

When creating the test suite and evaluating the different translation systems, the work process was as follows. First of all, we gathered the actual English data, with the goal of creating 300 segments. These range from single sentences to short paragraphs consisting of a few sentences. The majority of the data are real examples found through various online sources, although we have in some cases simplified them somewhat or added a bit of context to make it clear what sport is being discussed. A few of the sentences are purely synthetic.

When we were collecting the data, we picked out term candidates and assigned them translations. When finalizing the list of terms, one of the authors read through all the segments and inspected the term candidates with the goal of limiting internal discrepancy in the translations of a term that occurs more than one time in each sport. However, in order to reduce the impact of a single term on the overall outcome, he tried to limit the occurrences of a single term within a single sport to three. Even if a term together with its translation occurs three times in one sport, that term may also be found in the list within a different sport. An example is *penalty* which is a term in football and golf. While it can be translated as 'víti' in both sports, we also translate it as 'vítaspyrna' in football and 'refsing' in golf. 'Refsing' would not work as a translation for *penalty* in football and likewise, e.g., 'vítaspyrna' would not work as a translation of the word in golf (the head noun of the compound *vítaspyrna* is *spyrna* 'kick'). Furthermore, some terms occur as a noun and a verb within a single sport. An example is *foul*. We decided to count the occurrences of it as a noun separately from when it occurs as a verb, meaning that it may be found three times as a noun and three times as a verb in one and the same sport. Also, we treat the noun *foul* in basketball separately from, e.g., the term *technical foul*.

When the list of terms had been finalized, another author ran the data through automatic evaluation and prepared a document for two of the authors, different from the two already mentioned, for manual evaluation. The evaluation process is described and discussed in Section 4.

3.3 Problematic Translations

We evaluate translations generated by 34 MT systems. Out of the 34 systems, we received the original test suite segmentation from 15 systems, i.e. 300 segments with a 1-to-1 mapping to our original segments. The other 19 systems delivered more or fewer segments. For two systems, IR-MultiagentMT and ONLINE-B, all text for each sport was translated as one document and we recieved the translations as 5 lines, one for each document. For these documents we segmented the text using NLTK (Bird and Loper, 2004) and then used SentAlign (Steingrímsson et al., 2023) to match the translations to the correct segments. We fixed the other outputs mostly manually in order to be able to carry out our automatic evaluation approach. Table 3 lists the systems we fixed the output for, explains what was out of order and how we fixed it. In many cases the root of the problem seems to be the inability of the systems to translate long documents so they fail on multiple translations due to that.

4 Evaluation Methodology

4.1 Automatic Evaluation

We employ a rather simplistic approach to automatic evaluation. For each segment in our test suite, the English terms have been identified and their Icelandic counterparts registered.

System Name	Lines	Problem	Fix
Erlendur	300	Incorrect order.	Manually
IR-MultiagentMT	5	Documents returned as one line.	SentAlign
Mistral-Medium	302	Split segments.	Manually
ONLINE-B	5	Documents returned as one line.	SentAlign
UvA-MT	100	Many lines missing. Translated as documents but seems to stop after approx. 15 sentences in each document.	Manually
TowerPlus-72B	264	Many lines missing. Fails on long documents, gets stuck in a loop and starts repeating previous translations.	Manually
Qwen2-235B	221	Many lines missing for basketball.	Manually
Llama-4-Maverick	329	Many football segments split.	SentAlign
IRB-MT	593	Football starts repeating in a loop when 12 sentences are left, thus only contains 288 valid translations in total.	Manually
GemTrans	343	Football starts repeating itself when two lines are left, we remove all text produced after repetitions start.	Manually
Gemma-3-27B	301	Split segments.	Manually
Gemma-3-12B	303	Split segments.	Manually
EuroLLM-22B	199	All football in one line. (Translations in Swedish, not Icelandic.)	Manually
EuroLLM-9B	176	Gymnastics in one line. Football missing. (Translations in Swedish, not Icelandic.)	Manually
DeepSeek-V3	197	Basketball and football fails because the document is too long. Translations missing for these two sports.	Manually
CommandR7B	277	Multiple errors. Gymnastics don't finish but start repeating previous translations. Basketball translations phrased like headlines (and in German, not Icelandic).	Manually
CommandA	293	Some basketball translations missing.	Manually
AyaExpanse-32B	299	Missing translations.	Manually
AyaExpanse-8B	241	Many missing translations. Some translations repeated.	Manually

Table 3: The submissions by some of the MT systems did not contain the same number of lines as our test suite did. Section 3.3 describes how we tried to align the source to the correct translated segments.

System	Correct Terms	Chess	Basketball	Golf	Gymnastics	Football
Gemini-2.5-Pro	**54.57%**	**64.79%**	**52.26%**	51.61%	**58.93%**	**53.79%**
Shy	48.48%	61.97%	46.09%	50.00%	26.79%	51.03%
Erlendur	47.92%	39.44%	48.56%	46.77%	48.21%	49.66%
GPT-4.1	46.95%	43.66%	48.15%	46.77%	35.71%	48.97%
TranssionTranslate	43.35%	21.13%	45.27%	**54.84%**	25.00%	48.28%
★ ONLINE-B	43.07%	26.76%	42.39%	50.00%	25.00%	49.66%
TowerPlus-9B	42.94%	21.13%	45.27%	48.39%	28.57%	47.93%
hybrid	40.03%	26.76%	46.09%	19.35%	30.36%	44.48%
Claude-4	36.98%	18.31%	42.39%	35.48%	19.64%	40.69%
ONLINE-G	36.98%	22.54%	36.21%	40.32%	10.71%	45.52%
AMI	35.32%	14.08%	42.80%	43.55%	16.07%	36.21%
Gemma-3-27B	34.21%	19.72%	37.86%	37.10%	12.50%	38.28%
★ Llama-4-Maverick	32.41%	29.58%	38.68%	25.81%	7.14%	34.14%
NLLB	30.75%	12.68%	35.39%	19.35%	10.71%	37.59%
Mistral-Medium	30.06%	18.31%	31.28%	37.10%	8.93%	34.48%
■ GemTrans	27.01%	9.86%	33.33%	25.81%	7.14%	30.00%
SalamandraTA	24.38%	9.86%	28.40%	22.58%	14.29%	26.90%
Gemma-3-12B	23.41%	8.45%	27.98%	24.19%	7.14%	26.21%
■ IRB-MT	22.99%	8.45%	27.57%	25.81%	8.93%	24.83%
■ DeepSeek-V3	18.28%	22.54%	9.05%	25.81%	10.71%	24.83%
CommandA-MT	18.14%	4.23%	24.69%	16.13%	8.93%	18.28%
Llama-3.1-8B	14.40%	4.23%	17.28%	11.29%	0.00%	17.93%
■ CommandA	13.85%	5.63%	15.64%	12.90%	5.36%	16.21%
■ Qwen3-235B	12.88%	9.86%	3.70%	12.90%	5.36%	22.76%
■ TowerPlus-72B	9.14%	19.72%	1.65%	38.71%	12.50%	5.86%
■ UvA-MT	5.82%	7.04%	5.76%	16.13%	7.14%	3.10%
■ AyaExpanse-32B	3.74%	2.82%	5.35%	9.68%	0.00%	2.07%
Qwen2.5-7B	3.74%	0.00%	3.70%	4.84%	1.79%	4.83%
Mistral-7B	3.32%	0.00%	4.53%	6.45%	0.00%	3.10%
■ EuroLLM-22B	2.63%	0.00%	6.17%	6.45%	0.00%	0.00%
■ CommandR7B	1.66%	0.00%	0.41%	6.45%	0.00%	2.41%
★ IR-MultiagentMT	1.66%	11.27%	1.23%	0.00%	0.00%	0.34%
■ AyaExpanse-8B	1.39%	0.00%	2.47%	3.23%	0.00%	0.69%
■ EuroLLM-9B	1.25%	0.00%	2.88%	3.23%	0.00%	0.00%

Table 4: Automatic evaluation of each model across categories. The systems are in order of overall accuracy, with the highest scoring system being the only one that translated more than 50% of terms correctly. Accuracy is the ratio of correct translations to total term occurrences. The highest score for each domain is in bold letters.

We use these keywords when evaluating, by inspecting each translated segment and checking whether it contains the expected Icelandic term or terms. As Icelandic is an inflected language, we must consider all possible forms of the terms. For terms consisting only of one word, we look up all possible forms of the word in DIM, the Database of Icelandic Morphology (Bjarnadóttir et al., 2019), and if it is not found there, we manually list the forms. For multiword terms we manually create possible forms and list them. If any form of the term is found in the translation we count that as correct. We expect this approach to give us a close approximation of how correctly the MT systems translate the terminology. In some cases though, correct forms of terms may be missing, or a form is used that might make the translation ambiguous or wrong although our approach marks it as correct. There may also be variations of some terms, that we do not register but would be considered correct by a human judge. In order to inspect how close the automatic evaluation is to human judgments, we carry out a manual evaluation for comparison.

4.2 Manual Evaluation

For manual evaluation, we randomly sample segments from all five subdomains, depending on how many sentences they have in the test suite. For basketball and football we select 10 segments, 7 segments for chess and 5 for golf and gymnastics. We then collect translations for these segments from all evaluated systems. Human evaluators judge the translations and check if the term is correctly translated, without regard to the registered Icelandic term translations. Scores for each system are given as a percentage of correctly translated terms.

5 Results

We report on the results of our automatic evaluation approach and manual evaluation and compare the outcomes. In Section 3.3 we discuss how the outputs of some systems were problematic. In the results tables, we tag the systems that may be at a disadvantage due to other reasons than just their capability in translating from English into Icelandic. We use two tags: ■ for translations that had missing lines or repetitions, possibly because all segments for a given sport were being translated as one document, but the model failed to handle such long documents, and ★ for translations where we had to run sentence alignment to match translations to the original segments. These tagged systems may score lower than expected, when compared to the results in the general translation task (Kocmi et al., 2025b,a), in most cases likely due to inability to handle long documents. Two of the three systems that we realigned using SentAlign, ONLINE-B and Llama-4-Maverick, seem to score similarly to the systems in the general translation task, so splitting up and realigning may have had minimal effects in these cases.

5.1 Automatic Evaluation

We find that according to our automatic evaluation, see Table 4, only one system manages to translate over 50% of the terms correctly. This system, Gemini-2.5-Pro, also scores highest in all domains except for golf, where it has the second highest score. There is a markedly large difference between the highest scoring system and the next, but the systems in second to fourth place are not very far from each other.

It is also noteworthy that there can be a large difference between systems within domains, while the difference on average is not substantial. Only six systems translate more than 25% of the chess terms correctly, while half or more reach that threshold for basketball, football and golf. In spite of that two systems translate almost over 60% of the chess terms correctly.

We also looked at terms that occur twice or more in the test suite and inspected whether the ones that were translated correctly at least once were consistently translated correctly. Table 5 shows that this was rarely the case, with only two systems consistently translating over 50% of these terms correctly. This is a disappointment, as it indicates that even when a system translates the terminology correctly, it may be incidental.

5.2 Manual Evaluation

The results of the manual evaluation are given in Table 6. While there is a general consensus between the automatic and manual evaluation with respect to the order of best systems, there are some variations. The highest scoring system in the automatic evaluation, by a good margin, switches seat with the second best system in the manual evaluation. This being said, the sample was much smaller in the manual evaluation, with only 77 terms being checked in 37 segments. The difference between the top systems is thus only one correct translation.

System	Some	All	Cons.
Gemini-2.5-Pro	108	56	51.85%
Shy	99	44	44.44%
Erlendur	96	44	45.83%
GPT-4.1	100	45	45.00%
TranssionTranslate	96	43	44.79%
★ ONLINE-B	96	42	43.75%
TowerPlus-9B	99	40	40.40%
hybrid	96	32	33.33%
Claude-4	81	42	51.85%
ONLINE-G	87	38	43.68%
AMI	90	31	34.44%
Gemma-3-27B	78	36	46.15%
★ Llama-4-Maverick	75	33	44.00%
NLLB	77	30	38.96%
Mistral-Medium	70	29	41.43%
■ GemTrans	66	20	30.30%
SalamandraTA	60	24	40.00%
Gemma-3-12B	56	19	33.93%
■ IRB-MT	58	18	31.03%
■ DeepSeek-V3	58	5	8.62%
CommandA-MT	42	17	40.48%
Llama-3.1-8B	41	10	24.39%
■ CommandA	33	11	33.33%
■ Qwen3-235B	38	10	26.32%
■ TowerPlus-72B	26	5	19.23%
■ UvA-MT	22	3	13.64%
■ AyaExpanse-32B	12	3	25.00%
Qwen2.5-7B	11	2	18.18%
Mistral-7B	8	2	25.00%
■ EuroLLM-22B	8	1	12.50%
■ CommandR7B	7	0	0.00%
★ IR-MultiagentMT	6	1	16.67%
■ AyaExpanse-8B	5	0	0.00%
■ EuroLLM-9B	3	1	33.33%

Table 5: Consistency in term translation. 142 unique terms are seen more than once in the test suite. The table shows how many of them are sometimes translated correctly, and how many of them are consistently translated correctly, every time they occur. Consistency is given as a percentage of terms consistently translated correctly of reoccurring terms that are translated correctly at least once.

System	Accuracy
Shy	77.92%
Gemini-2.5-Pro	76.62%
Erlendur	72.73%
GPT-4.1	71.43%
TowerPlus-9B	68.83%
TranssionTranslate	66.23%
hybrid	62.34%
★ ONLINE-B	62.34%
ONLINE-G	61.04%
Claude-4	55.84%
AMI	54.55%
Gemma-3-27B	42.86%
★ Llama-4-Maverick	40.26%
■ GemTrans	37.66%
Mistral-Medium	35.06%
NLLB	29.87%
■ IRB-MT	28.57%
■ DeepSeek-V3	28.57%
Gemma-3-12B	27.27%
SalamandraTA	23.38%
■ CommandA-MT	19.48%
■ TowerPlus-72B	18.18%
Llama-3.1-8B	11.69%
■ Qwen3-235B	11.69%
CommandA	11.69%
■ UvA-MT	6.49%
■ AyaExpanse-32B	3.90%
★ IR-MultiagentMT	2.60%
Mistral-7B	1.30%
■ EuroLLM-22B	1.30%
Qwen2.5-7B	0.00%
■ AyaExpanse-8B	0.00%
■ CommandR7B	0.00%
■ EuroLLM-9B	0.00%

Table 6: Manual evaluation of a sample of translations from each system.

The accuracy in the manual evaluation is higher than in the automatic one. This could be expected as it is likely that some word forms or variations of the terms are not registered in our lists even though a human would consider them correct and mark them as such when evaluating them. Even though the manual evaluation paints a prettier picture in this regard, it shows, like the automatic evaluation, that even the best systems are still failing quite often when translating this specialized, but common, vocabulary.

6 Limitations

There are various limitations to the current work, some of which are discussed below.

In various segments, the English terms are a part of a larger compound. We have tried to avoid including such compounds as listed terms in some places but where we think it is appropriate, we try

to make the translated terms capture this fact. An example is the term *point guard*, which is translated as 'leikstjórnandi, ás'. However, in one segment, *point guard* is part of the compound ***star** point guard*.[5] In addition to the translations 'leikstjórnandi, ás', we include '**stjörnu**leikstjórnandi, **stjörnu**ás'. When translating to Icelandic, English compounds can sometimes be reworded as a phrase rather than a single compound. An example is *Liverpool supporter* which appears in one of the segments in our football set. *Supporter* is a term in the set and one of the translations we give for it is 'stuðningsmaður'. By rephrasing we could translate *Liverpool supporter* as 'stuðningsmaður Liverpool', meaning 'a supporter of Liverpool', but if we would insist on translating it as a compound, we could translate it as 'Liverpool-stuðningsmaður' — and we give that as a possibility in our test suite in that particular segment. In some cases, however, we may have failed to notice that a term is a part of a compound but when a translation system translates the whole term as "one" word (without spaces or without rephrasing) the automatic evaluation would mark it as incorrect. A dataset with more synthetic data could more easily avoid such compounds.

The manual evaluation makes it clear that we miss out various translation possibilities we did not think of, showing that the automatic evaluation is limited. Furthermore, only one annotator evaluated each translation. If multiple annotators would have evaluated the same translations they would possibly have had some disagreements. That would also have allowed us to calculate inter-annotator agreement, which could give us an indication of whether a manual evaluation such as the one we carried out is straightforward or whether it demands multiple annotators for each segment.

We try to reflect actual use in our translations. That is not an easy task, especially when term use in written texts, such as in fairly formal news coverage, differs from spoken language use. An example might be *layup* in basketball which we only translate as 'sniðskot'. However, searching for a string starting with "layup" in the Icelandic Gigaword Corpus gives 219 results. 'Sniðskot' has a formal feel to it but it captures the English term. Sometimes, however, we were not sure whether the Icelandic translation captures it completely or is well-known enough. An example is the basketball term *screen*, which occurs as a noun in the test suite and is translated as 'hindrun' in the Icelandic version of the FIBA basketball rule book;[6] we made the decision to give an Icelandic spelling version of *screen*, i.e., 'skrín', in addition to 'hindrun' which goes back centuries in the language and can be translated as 'hindrance, obstacle', even though we translated *layup* as 'sniðskot' only (a relatively new compound in the language whose head is *skot*, meaning 'shot'). Further work that takes a closer look at the actual usage would be interesting and working with professionals in each sport on the translations of the terms would be ideal.

Which words should be given as terms in a work like this can be debated. We have the verb and the noun *win* in various places listed as terms but as winning is not specifically found in one sport but not the other we might want to exclude some such words when we are exploring a translation system's ability to translate the vocabulary in a certain branch of sport. The same goes for, e.g., parts of the body: A word like *shin* is probably used more in a sport like football than golf but, nevertheless, one could certainly disagree with our decision of leaving that in as a football term in our dataset.

As natural data are the bulk of segments in the test suite, it is not always clear from the segment's context alone what the sport in question is. In a few places, we added information in the segments. Instead of using the original unchanged in the segment *A screen is the legal action of a player who, without causing undue contact, delays or prevents an opponent from reaching a desired position* we added "In basketball" to make it clear what the sport is: ***In basketball**, a screen is the legal action of a player who, without causing undue contact, delays or prevents an opponent from reaching a desired position*. However, we did not generally focus on this when finalizing the dataset and what impact this has on the translation scores merits further study. Moreover, a synthetic dataset could control for this.

[5] Note that although the term *point guard* is a compound on its own, we focus here on *star* + *point guard*, where only *point guard* is a term according to our dataset and not the whole compound *star point guard*.

[6] See, e.g., the 2017 version here: https://www.kki.is/library/Skrar/Leikreglur_i_Korfuknattleik_2018.pdf.

7 Conclusions and Future Work

Overall, the results indicate that more attention needs to be paid to the language of sports in LLMs and MT and the generally low scores confirmed our suspicion that MT systems have a hard time at translating sports texts adequately. Even the highest scoring systems do not do a very good job at translating the language of sports, and in all cases consistency is lacking.

Categorizing the errors could be useful for analyzing comparative differences between system types. Such categorization could also be useful for building MT systems better suited for translating sports language.

We intend to use our test suite to automatically evaluate new systems and keep track of their competence in translating sports terminology. While the sports we selected are some of the most popular ones that have a specialized vocabulary in Icelandic, expanding the test suite by adding more sports could be useful. Covering the terminology of each sport more thoroughly could help us get more accurate results and having at least two occurrences of each term could make our consistency evaluation more precise. It would also be interesting to translate the keyword list into more languages than Icelandic to investigate whether this problem is specific to Icelandic.

Acknowledgments

This project was funded by the Language Technology Programme for Icelandic 2024–2026. The programme, which is managed and coordinated by Almannarómur, is funded by the Icelandic Ministry of Culture, Innovation and Higher Education.

We thank Árni Jóhannsson, Eiríkur Stefán Ásgeirsson and Jóhannes Gísli Jónsson for answering questions on specific terms relating to basketball, golf and chess, respectively. Thank you to Ágústa Þorbergsdóttir for discussions on terminology work in Iceland. We would also like to thank two anonymous reviewers for valuable feedback on the paper.

References

Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinþór Steingrímsson. 2024. Killing Two Flies with One Stone: An Attempt to Break LLMs Using English-Icelandic Idioms and Proper Names. In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458. Association for Computational Linguistics.

Starkaður Barkarson, Steinþór Steingrímsson, and Hildur Hafsteinsdóttir. 2022. Evolving Large Text Corpora: Four Versions of the Icelandic Gigaword Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2371–2381. European Language Resources Association.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217. Association for Computational Linguistics.

Kristín Bjarnadóttir, Kristín Ingibjörg Hlynsdóttir, and Steinþór Steingrímsson. 2019. DIM: The Database of Icelandic Morphology. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 146–154. Linköping University Electronic Press.

Lise Lotte Weilgaard Christensen, Hanne Erdman Thomsen, Bodil Nistrup Madsen, Anna-Lena Bucher, Henrik Nilsson, Claudia Dobrina, Håvard Hjulstad, Åsa Holmér, Johan Myking, Anita Nuopponen, Sirpa Suhonen, Anu Ylisalmi, and Ágústa Þorbergsdóttir. 2025. The Nordic Terminology Community. Research and practice. In *Terminology throughout History. A discipline in the making*, pages 327–364. John Benjamins.

Knattspyrnulög. 1907. Íþróttafjelag Reykjavíkur.

Steinunn Rut Friðriksdóttir. 2024. The GenderQueer Test Suite. In *Proceedings of the Ninth Conference on Machine Translation*, pages 327–340. Association for Computational Linguistics.

Örn Hrafnkelsson and Jökull Sævarsson. 2014. Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books. In *Language Resources and Technologies for Processing and Linking Historical Documents and Archives Deploying Linked Open Data in Cultural Heritage – LRT4HDA, LREC 2014*. European Language Resources Association.

Atli Jasonarson. 2025. Áhliða óvinamegin: Um orðaforða Knattspyrnulaga frá 1907. Paper presented at the 38th Rask-ráðstefna um íslenskt mál og almenna málfræði, January 24th, University of Iceland.

Rebecca Knowles, Samuel Larkin, Marc Tessier, and Michel Simard. 2023. Terminology in Neural Machine Translation: A Case Study of the Canadian Hansard. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 481–488. European Association for Machine Translation.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda,

Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671. Association for Computational Linguistics.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and Scalable Sentence Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263. Association for Computational Linguistics.

Fine-Grained Evaluation of English-Russian MT in 2025: Linguistic Challenges Mirroring Human Translator Training

Shushen Manakhimova[1], Maria Kulinovskaya[2], Ekaterina Lapshinova-Koltunski[3], Eleftherios Avramidis[1],
[1]German Research Center for Artificial Intelligence (DFKI)
firstname.lastname@dfki.de
[2]Saarland University, maria.kunilovskaya@uni-saarland.de
[3]University of Hildesheim, lapshinovakoltun@uni-hildesheim.de

Abstract

We analyze how English–Russian machine translation (MT) systems submitted to WMT25 perform on linguistically challenging translation tasks, similar to problems used in university professional translator training. We assessed the ten top-performing systems using a fine-grained test suite containing 465 manually devised test items, which cover 55 lexical, grammatical, and discourse phenomena, in 13 categories. By applying pass/fail rules with human adjudication and micro/macro aggregates, we observe three performance tiers. Compared with the official WMT25 ranking, our ranking broadly aligns but reveals notable shifts.

Our findings show that in 2025, even top-performing MT systems still struggle with translation problems that require deep understanding and rephrasing, much like human novices do. The best systems exhibit creativity and can be very good at handling such challenges, often producing more natural translations rather than producing word-for-word renditions. However, persistent structural and lexical problems remain: literal word order carry-overs, misused verb forms, and rigid phrase translations were common, mirroring errors typically seen in beginner translator assignments.

1 Introduction

Unlike standard test sets, which consist of randomly selected source material, a test suite contains 'extra-credit' problems: the source items, deliberately designed to be challenging in translation, similar to the 'rich points' method (Nord, 1997) or preselected items method (Egdom et al., 2019) in translator training. These focused items are often lexical units or grammatical structures requiring non-literal approaches based on deep situational understanding and coherent target-language encoding. Successful translations are expected to abstract away from source form, displaying both familiarity with standard translation techniques and creativity in adapting them to individual contexts. While isomorphic translations may be formally grammatical, they are typically judged suboptimal. In translation didactics, such tasks test translators' ability to identify and resolve translation problems.

The fine-grained linguistic test suite has been partially in development over the last few years (Macketanz et al., 2022; Manakhimova et al., 2023, 2024).[1] Its original purpose was to track MT systems' ability to handle specific source language phenomena, with evaluation strategies focused on translating those targeted items. This approach provides valuable, fine-grained insights into MT systems, highlighting system strengths across a wide range of phenomena and establishing objective grounds for comparison. However, the recent dominance of large language models (LLMs) in MT has prompted us to reconceptualize the test suite's role. Beyond its original function of testing specific linguistic categories, we also view it as a **diagnostic tool for identifying overarching translation challenges** that persist across current English–Russian MT systems. This evolution reflects the need to understand systemic issues in MT. Accordingly, we have tightened **our quality requirements to match those of a university professional translation training**. Our evaluation focus has shifted from assessing the handling of specific linguistic categories to evaluating overall translation quality when processing these 'extra-credit' problems. We no longer make concessions for 'gist translation quality,' instead expecting the publication-standard quality of human translation as defined by Ahrenberg (2017, p. 21). This refined approach maintains our ability to compare systems based on the ratio of successfully handled

[1]The wider project also maintains suites for other directions (e.g., German-English, English–German, English–Portuguese in Avramidis et al., 2020; Macketanz et al., 2021; Avelino et al., 2022), which we do not analyze here.

Proceedings of the Tenth Conference on Machine Translation, pages 866–877
November 8-9, 2025

items across the entire test suite, while simultaneously revealing persistent issues and blind spots that transcend formal categories. We supplement our qualitative analysis with statistical results and system rankings, comparing these findings with WMT's official human evaluation.

2 Method

2.1 Test suite overview

For the English–Russian part of the test suite, the 13 evaluation categories cover a broad range of translation challenges. They include ambiguity, collocations, compounds, false friends, and multi-word expressions, which test lexical precision. Morphosyntactic control is addressed through case government, function words, passive voice, subordination, and verb valency. The suite also targets discourse as well as stylistic aspects via personal pronoun coreference and onomatopoeia. Together, these categories span the main lexical, grammatical, and pragmatic difficulties that can arise when translating from English to Russian.

Test items consist of one (or occasionally more) source sentence(s) plus an associated set of evaluation rules. These rules comprise hand-crafted regular expressions and fixed strings of translation outputs. Test items are either created manually by linguists or sourced from existing corpora and curated for the target phenomenon. An ideal example for a test suite should require the interpretation of the message and deep restructuring in the target language without being vague or context-dependent. Such examples also give rise to greater variability in translation, and can be identified as having a higher entropy of translation solutions, a known measure of the source item difficulty in translation (Carl and Schaeffer, 2017; Wei, 2022; Kunilovskaya et al., 2025). Examples from the *Resultative* subcategory (category: Verb valency) seem to represent true challenges in the English–Russian translation. For example, *The skiers skied the trail clean of snow.*

As MT technology has shifted from phrase-based systems to NMT and now LLMs, error profiles have also changed. The suite has therefore been revised over time: we have added phenomena, increased item counts, and introduced longer or structurally richer sentences to stress contemporary systems. MT outputs, after being evaluated by the test suite, have also been utilized to create challenge sets for WMT metrics (Avramidis et al., 2023, 2024). This year's revision of the test suite excluded several items that either misrepresented their category or lacked context independence to be fairly evaluated.

Given the increased variation in translation solutions offered by the submitted systems, the automatic pass-fail rules and annotation guidelines were tightened this year to reflect the requirements that would be applied to translations considered as part of a university professional translator training.

2.2 Scoring

The evaluation results for the categorized items are produced semi-automatically: hand-written regular expressions as well as fixed strings (correct and incorrect translations from earlier MT system outputs) capture expected *correct* and *incorrect* renderings, and any remaining cases are adjudicated by a linguist. Regular-expression design leverages prior experience with MT outputs and aims to maximize coverage; however, novel outputs routinely require human judgment.

Since this evaluation aims to compare the systems fairly, only the test items that have a valid judgment for all systems are included in the calculation. If a test item has a judgment neither by regular expressions nor by the annotators for any MT systems, we exclude it from the calculation. As a result, not all test items that had originally been designed can be used in our calculations.

For system comparison, we first identify the highest-scoring system and then test all others against it using a one-tailed Z-test with $\alpha = 0.95$. Systems not significantly worse than the top system form the first performance cluster; we mark the best systems in bold in the result tables. Because categories and phenomena (subcategories) differ in size, we report three complementary aggregates: micro-average (accuracy over all items, item-weighted), category macro-average (mean of category-level accuracies), and phenomenon macro-average (mean of phenomenon-level accuracies).

2.3 Manual annotation procedure

This year marks the third evaluation of English–Russian systems using our test suite. As in previous years, manual intervention was necessary to process system outputs that could not be automatically evaluated. At the beginning of this year's evaluation, this referred to 44.83% of outputs on average. Three annotators divided the workload,

with each responsible for a disjoint subset of the data.

We did not compute inter-annotator agreement (IAA), as the evaluation workflow did not involve multiple annotators independently labeling the same outputs. Instead of computing IAA, we relied on the extensive, linguistically motivated rule set refined over several years. These rules served as a shared reference, reducing subjectivity; borderline cases were resolved in group discussion.

Annotators were instructed to focus on the targeted source-language phenomenon, provided that a translation candidate meets basic standards of accuracy and fluency. Above all, the translation should faithfully convey the original message while adhering to the norms and conventions of the target language.

The evaluation is relative: in disputable cases, comparing translation candidates helps determine what is achievable for a given item. In professional translator training, it is common to assess solutions against available or hypothesized alternatives. For example, Bittner (2020, p. 172) observes: *"Good translation quality can only be better translation quality, just as bad translation quality can only be worse translation quality. There is no use dismissing a translation solution as unacceptable unless a better alternative can be produced."* At the same time, it is possible that none of the proposed solutions is acceptable, or that two solutions using different techniques are equally valid.

When a source item is aimed at evaluating more than one phenomenon, a translation is considered correct even if some parts are suboptimal (but not unacceptable). This ensures that effective strategies and creative handling of the target phenomenon are recognized. To illustrate the kinds of defects that were tolerated, consider Example (1), where the source is categorized as *Modifying Comparison* and both translations are accepted as correct. The second version, however, is preferable: it omits the possessive pronoun её ("her") in the subordinate clause, avoids semantic tautology in rendering *expertise*, and dispenses with the redundant demonstrative pronoun того. Importantly, both variants successfully address the lexical challenges that weaker systems often mishandle. For instance, many systems reproduced the English collocation *extensive level of expertise* as обширный уровень экспертизы.

(1) Her level of expertise was not as extensive as her employer had hoped.

a. Уровень её профессиональной квалификации оказался не таким высоким, как того ожидал её работодатель.

b. Уровень её квалификации оказался не таким высоким, как надеялся работодатель.

In cases where the source was ambiguous or difficult to interpret (e.g., *he ran under the porch*), a translation was judged correct if it provided a plausible and contextually logical reading consistent with real-world knowledge. For example, он (по)бежал под крыльцо for the source above remains questionable.

Although the items are designed to be context-independent, some may admit multiple interpretations in a broader context. In such cases, evaluation favors the most prototypical or expected reading, while unusual or exotic contexts are disregarded. Finally, translations that ignore potential of the target language for optimal information packaging – often requiring creative reconceptualization of the message – and instead follow the source language structures in a routine, linear manner, were not accepted. They might be grammatically correct, but in dissonance with the conventional usage of the target language.

2.4 Experiment Setup

Although the full test suite was applied to 42 systems submitted to the WMT25 Shared Task, this paper reports statistical comparisons for a representative subset of 10 systems. The selection is based on the official Error Span Annotation

3 Results

3.1 System Performance Overview

This section reports system-level performance (overview and hierarchy) and category-level difficulty, following the scoring protocol in Section 2.2 Aggregate accuracies per system and per category/phenomenon are provided in the Appendix tables, along with the test suite system ranking and the official WMT25 ranking (Kocmi et al., 2025b).

The performance distribution reveals three distinct tiers: high performers Wenyiil (Wang, 2025), Algharb (Xu, 2025), Yandex (Karpachev et al., 2025), and Gemini (Finkelstein et al., 2025);

mid-tier systems Claude-4, DeepSeek-V3, GPT-4.1, and Shy-hunyuan-MT (Zheng et al., 2025, 89.4–91.9%); and lower-tier systems CommandA (Kocmi et al., 2025a) and UvA-MT (Wu et al., 2025, 83.2–88.4%).

Compared to the WMT25 official ESA ranking, Yandex moves from the 8–10 cluster into the top cluster. Conversely, Shy-hunyuan-MT-hunyuan-MT, ranked 2 under ESA, falls into the middle cluster in our suite, Claude-4 and GPT-4.1 also belong to the middle cluster; both exhibit the same weakness on verb semantics (71.4%), and GPT-4.1 additionally scores low on long-distance dependencies and interrogatives (81.5%).

Peformance of Constrained vs. unconstrained models. In our test suite, constrained MT systems frequently occupy the top cluster, consistent with scoring that rewards precise handling of hard, localized phenomena and conservative choices under ambiguity. In the WMT25 official ranking, however, unconstrained systems rise, reflecting strengths in fluency, stylistic naturalness, and document-level coherence. Systems that bridge both profiles narrow the gap between the two.

3.2 Category Difficulty Analysis

MWE represents the most challenging category with only 80.4% average accuracy across all systems, followed by *Verb semantics* (81.4%) and *Verb valency* (87.4%). These categories demonstrate the complexity of handling idiomatic expressions and verb–argument structures in English–Russian translation. Conversely, *Lexical Ambiguity* proves easiest (97.0% average), with eight systems achieving perfect scores, indicating strong disambiguation capabilities across translation systems. *Function words* and *Subordination* also show high accuracy (93.6% and 94.3% respectively), suggesting robust handling of grammatical structures.

Verb semantics exhibits the largest performance gap (57.1%) between systems, with Wenyiil achieving 100% while UvAMT manages only 42.9%.

3.3 Linguistic Analysis

In this section, we summarize the overall patterns and translation strategies revealed in the manual analysis, along with some notable peculiarities of individual systems. We then turn to the most persistent challenges: (a) difficulties rooted in English source structures and (b) recurring problems with Russian target-language conventions that MT systems struggle to master. None of the highlighted issues is ubiquitous; for each example, we provide a more acceptable version drawn from the available translations. Translations marked with an asterisk are considered suboptimal. Generally, a comparison with previous years' submissions indicates noticeable improvements across most problem areas.

Overall translation patterns. This paragraph offers some high-level observations from annotating the 2025 submissions in comparison with previous years. We note an improvement in the variation of generated output, suggesting greater creativity and a stronger ability to recast the original message in new forms, rather than reproducing the formal and semantic structures of the source language.

Translations from the strongest 2025 systems demonstrate an increased capacity to do what the test suite examples force them to do: they **move beyond literal translation** and re-package the original message into a form that is natural in the target language. For a human translator, accomplishing this requires careful extraction of the intended meaning, imagining the described real-world situation, and expressing it in a way that aligns with conventional norms and expectations. The distinction between a literal strategy and a more interpretative approach is illustrated in Example (2). The example highlights the ability of the system to infer the contextual meaning of the descriptive verb "shuddered" and generate a plausible rendering of the situation.

(2) They shuddered home under the hailstorm.
   a. Они брели домой под градом, ежась от холода.
   b. *Они дрожащими вернулись домой под градом.

Re-creating a situation in another language may require modifying the set of properties by which it is recognized in the target language. In 2025, automatic translation systems are better at adding necessary elements and omitting redundant ones. For instance, *he's a fabulous inspiration* is rendered as он потрясающий <источник> вдохновения, while *Many people are concerned about High Street* becomes Многих беспокоит <состояние> главной улицы.

The ability of a system to take context into account and coordinate elements into a coherent

whole can also be observed at the sentence level. The source in Example (3) evokes a snapshot-like scene. Re-creating this scene in Russian requires abandoning the English mode of depiction and adopting a different strategy in the second clause. The asterisked translation illustrates a common problem: mismatched aspect forms in coordinated verbs, which disrupts the natural flow of information.

(3) Paula entered the small souvenir shop and took her time browsing through the magazines.
   a. Паула зашла в небольшой сувенирный магазин и принялась рассматривать журналы.
   b. *Паула вошла в небольшой сувенирный магазин и не спеша просматривала журналы.

This enhanced interpretative capacity reflects greater sensitivity to the functional potential of expressive means and the ability to deploy alternative but appropriate forms in the target language. As illustrated in Example (4), Algharb recognizes that paired synonyms are typical in English but generally avoided in Russian, while in Example (5), Algharb creatively re-packages the information to arrive at a conventional Russian rendition.

(4) Despite the neat and tidy ending to Season 3 . . .
   a. Несмотря на вполне законченную концовку третьего сезона, . . .

(5) . . . there was a delivery charge on top.
   a. . . . к сумме добавилась плата за доставку.

These examples are presented to highlight some of the advances in the technology. At the same time, they do not provide a complete picture, as persistent problems remain.

Where possible, the same systems resort to suboptimal **crude unpacking of English secondary predicates into full clauses**, producing wordy, redundant, and clumsy (but not ungrammatical) sentences. This strategy disrupts the information flow: in discourse, each full predicate conventionally signals a step forward in the narrative. By upgrading secondary predicates to main clauses, the translation introduces artificial shifts in topic–comment structure, resulting in unmotivated changes of focus and sentences that can appear contradictory or unclear. It is not uncommon that this tendency is coupled with a known redundancy of functional words such as auxiliaries, pronouns, and connectives (esp. consecutive connectives перед тем как, после того как).

Newer systems increasingly prefer **gender-neutral realization in the cases where gender is not explicit**, e.g. in generic contexts or in first person in Russian outputs. In contrast to English, Russian has explicit grammatical gender marking not only on pronouns, but also on past verb forms, participles, and adjectives. They are congruent with subjects. In contrast to previous years, systems for the first time consistently produced *explicit inclusive* forms, i.e., forms that include both masculine and feminine forms: купил(а) (“bought (m/f)”), мог(ла) (“could (m/f)”), готов(а) (“ready (m/f)”). The pattern is consistent across 33 systems (out of 42 evaluated systems) and various categories, indicating that newer models increasingly prefer gender-neutral realizations. Additionally, some systems (GPT-4.1, Algharb, Gemini) demonstrate the use of gendered profession names where the associated person is female (учительница, предпринимательница).

When faced with a **faulty input**, some models return a translation of the more plausible corrected source version, like a human translator would do. This is counteracting the automated tendency known as ‘garbage in, garbage out’.

Source challenges and target issues. In this part of the analysis, we first describe recurrent and problematic translation patterns in Russian that are triggered by specific source-language items. We then turn to a second group: target-language categories that consistently prove difficult for MT systems regardless the source.

(a) Source phenomena as error triggers. The most prominent defects in automatic translation stem from the literal transfer of source-language lexical and grammatical features. In particular, we highlight issues arising from (i) reproducing the word order of the source sentence, (ii) neglecting the contrastive use of pronouns and connectives, and (iii) calquing lexical frequency patterns and collocations. These problems complement those discussed in Section 3.2.

Unlike English, Russian relies heavily on **word order** to structure information. The most important, focused information typically occurs at the end of a

sentence, whereas English allows more flexibility; the sentence-final position in English can be filled with adverbials of time and place, prepositional objects, and other elements. Failure to identify the focused element in the English source and promote it to the sentence-final position in Russian, therefore, disrupts the natural flow of information, even in isolated sentences. A typical case is presented in Example (6). The topical sentence member им ("them") is awkward at the end of the Russian sentence in (6-b).

(6) When students walk into our classrooms, the course objectives are given to them right up front.
   a. Когда студенты заходят в наши аудитории, им сразу же сообщают цели курса.
   b. *Когда студенты заходят в наши аудитории, цели курса сразу же сообщаются <им>.

Misalignment in information structure is particularly noticeable in cleft sentences (e.g., *It wasn't until ...*) and elliptical constructions (e.g., *She asked the kids to stay, and the adults too; Laura drank the milk last night, or perhaps the juice; I met Aisha yesterday, but not her daughter*). As illustrated in Example (7), failing to place emphasized information at the end produces sentences like in (7-b) that are immediately recognizable as translations.

(7) After all it was not war that completely ravaged East Asian states in 1997.
   a. В конце концов, в 1997 году государства Восточной Азии разорила вовсе не война.
   b. *В конце концов, это была не война, которая полностью опустошила восточноазиатские государства в 1997 году.

Automatic translations into Russian often **overuse possessive pronouns**, mirroring their higher frequency in English. Example (8) shows sentences where *their* is rendered as <их> in Russian. While grammatically correct, these translations add possessive markers that a human translator would likely omit, resulting in a style that is formally acceptable but less natural (see also Example (1)).

(8) Despite <their> intense feelings for one another, it seems as though the two heroes might never remain together.
   a. Несмотря на <их> сильные чувства друг к другу, кажется, что этим двум героям никогда не быть вместе.

Indefinite pronouns (someone, anywhere, every, all) also contribute to a significant level of disfluency in machine translation. The apparent one-to-one correspondences have different usage patterns and frequencies (every $\neq$ каждый, all $\neq$ все).

Finally, **lexical problems** – the choice of words, collocation and idioms – are as pervasive as structural difficulties. Occasionally, many systems would find a particular word in the source language difficult and fall victim to literal translation, false friends, undetected idioms or terms. Example (9) shows a typical MT output, where sponsor is translated as спонсор. At the same time, in Russian, this word and derivatives from the same root rarely carry the "legislative initiator" meaning.

(9) They persuaded Kennedy and some other Senator to jointly sponsor the legislation, but I can't remember which one.
   a. *Они убедили Кеннеди и ещё одного сенатора совместно выступить спонсорами законопроекта, но я не помню, кого именно.

The hallmark of low-quality machine translation is translating every occurrence of *enjoy* with наслаждаться, and *people* with люди, to give examples of typical frequency calques seen in the analysis of this year.

(b) Target phenomena as persistent difficulties. A number of Russian categories can be problematic because they are not directly marked in the source but are obligatory in Russian. These categories are known to be difficult for Russian language learners, too. For example, English often encodes *verbal aspect* (a grammatical category that characterizes an action with regard to its internal temporal structure, such as whether it is ongoing, completed, repeated, or habitual) through contextual or grammatical means, while Russian uses a *lexico-grammatical system* (perfective, imperfective verbs). The translator is compelled to make a lexical choice that cannot be carried over from the source, since the category is not explicitly expressed there. Instead, the decision must be guided

by contextual cues and world knowledge, with the challenge lying in correctly reading those signals.

Verbal aspect: Automatic translation into Russian often struggles with maintaining aspectual coherence when rendering coordinated English verbs in the past tense. In Example (10), the first translation maintains consistent temporal and aspectual framing by using two perfective verbs (раздал and получил). By contrast, the incorrect version (10-b) uses an imperfective verb (раздавал) in the first clause. This creates a mismatch with the following perfective verb, producing an incoherent sequence: the first action is presented as ongoing or habitual, while the second is punctual and bounded. The result is an aspectual clash that disrupts the event structure of the sentence.

(10) The teacher handed out worksheets, but I didn't get one.
   a. Учитель раздал рабочие листы, но мне не досталось.
   b. *Учитель раздавал рабочие листы, но я их не получил.

In Example (11), the source communicates a polite encouragement and requires an imperfective verb (снимать) used in (11-a). Translations with possessive pronouns and perfective verb (снять) are suboptimal, because they sound like a command or instruction like in (11-b).

(11) Do take your coat off.
   a. Снимайте пальто (прошу вас).
   b. *Снимите своё пальто.

Another interesting example related to aspect is given in (12). The verb разъедайтесь used in (12-b) is an imperfective form of the verb разъесться, one of the meanings of which is 'to enjoy'. However, this verb is never used in the imperfective form with this meaning. Moreover, the given example contains an idiomatic expression which should not be translated literally word by word. Instead, a corresponding idiomatic expression should be used in Russian, as in (12-a).

(12) Enjoy your meal.
   a. Приятного аппетита.
   b. *Разъедайтесь обедом.

Beyond that, this output also contained the explanation by the system: *Note "Enjoy your meal" can be translated more literally as "Eat your meal with pleasure", but* Разъедайтесь обедом *is a more common colloquial way to say it in Russian*. This explanation is wrong.

Nominalisations: One important component of human translator training is drawing attention to linguistic categories that tend to be underrepresented in translation. Based on our analysis, in MT, these include *nominalizations* and *ellipsis*.

A variety of English subordinate clauses can be rendered as nominal phrases in Russian. This strategy helps avoid unnecessary nesting and reduces sequences of functional words, such as в том, что; это что-то, над чем; до тех пор, пока, resulting in a text that reads more naturally in Russian. In Example (13), the subordinate clause *he's retired* is rendered as the phrase после завершения карьеры in the accepted translation, whereas a weaker system (13-b) fails to apply this transformation.

(13) As previously documented, he discussed what his next move will be now that he's retired from in-ring competition.
   a. Как уже сообщалось ранее, он рассказал о своих дальнейших планах после завершения карьеры рестлера.
   b. *Как было задокументировано ранее, он обсудил, каким будет его следующий шаг теперь, когда он завершил карьеру активного борца.

Finally, a translated text can often be recognized by its unnaturally complete structure, with elements such as pronouns, copula verbs, and connectives explicitly spelled out where they could easily be inferred from the context. In other words, translated language underuses ellipsis.

There is a clear distinction between weaker and stronger MT systems in this regard. Stronger systems are less likely to produce a subject in subsequent clauses when it is identical to the subject of the main clause, which aligns with natural Russian usage and enhances fluency. For example, in Мы сделаем А, как только мы получим Б ("We shall do A, as soon as we get B"), the second мы would normally be omitted. Similarly, the copula verb in sentences such as Он заметил, что она <была> печальна ("He noticed that she was sad") should be omitted; however, failure to follow this pattern is pervasive even among strong

systems. These issues rarely impede understanding and are generally tolerated in evaluation, but they signal a lack of expected quality and function as an indicator of professional translation proficiency.

Due to space constraints, this description includes only the more pervasive defects of MT. Other flaws, which we want to flag, include contrastive connectives (esp. but translated as но where а is required), confusion caused by epistemic *would* and *could*, the limited use of short adjectives in predicative function, failures to build adverbial participial clauses as required by Russian school grammar, etc.

4 Discussion: Lessons Learned

The test suite was revised to exclude items with questionable categorization or insufficient context-independence for translation in isolation. The rules have become less permissive in terms of fluency and accuracy.

There are some similarities between the translation patterns produced by NMT and by learner translators (see the detailed analysis in Kunilovskaya et al., 2023). These similarities are most visible in source-language–triggered issues, such as the placement of adverbials and prepositional objects, or the preference for analytical (instead of synthetic) forms of future and passive verbs. This points to the phenomenon of shining through (as defined by Teich, 2003), which belongs to the phenomena of translationese (Gellerstam, 1986), i.e., specific features of translated language that make it different from non-translated original language production. However, compared to human translators, NMT systems (especially those outside the top tier) more often generate sentences that obscure the message, overcomplicate structure, or introduce redundancy. Such output departs from target-language conventions in ways that sound recognizably non-human, often failing to produce a coherent text that conveys a clear, plausible situation.

We observed that the same systems might pursue different strategies depending on the conditions. Faced with the absence of isomorphic structures in the target language, they are capable of impressively creative solutions. In less demanding contexts, however, they tend to revert to routine, near-literal strategies that overlook the target language's potential for more optimal expression.

5 Conclusion

In 2025, state-of-the-art English–Russian MT presented by the latest LLM-based systems shows substantial progress in performance, yet it continues to display important weaknesses. Our fine-grained evaluation revealed that even top-performing models still falter on translations requiring deep comprehension and nuanced rephrasing, much like human novice translators. At the same time, the best systems exhibit marked improvements on such 'extra-credit' items by re-structuring and wording translations in a more natural Russian style instead of relying on word-for-word renditions. This shift toward non-literal, context-aware translation indicates that current MT can approximate some of the flexible strategies employed by skilled human translators. Notably, general-purpose LLMs (e.g., GPT-4.1, Claude-4) only attained mid-tier accuracy on our test suite, underscoring that even massive generalist models have not fully solved certain linguistic subtleties Specialized MT engines thus continue to hold an edge on fine-grained challenges, though the gap is beginning to narrow as new models adopt more human-like problem-solving approaches.

Limitations

A limitation of our current evaluation design is its reliance on a binary correctness—each test item is marked as either correct or incorrect based on regular-expression matching or manual adjudication. While this design facilitates scoring and result aggregation, it inevitably lacks the granularity needed for a more nuanced evaluation of translation quality, especially when annotators are faced with a human-like variation of MT outputs for less straightforward examples.

The second most notable limitation is the sentence-level nature of examples, which provides a reduced opportunity to track translation problems that might arise from the discourse level. It is not clear whether MT models would employ sentence splitting and merging as well as redistribution of semantics across several sentences if they were faced with larger spans of text to operate on.

Next, the test suite requires further revision to strengthen its construct validity. In particular, source items should foreground the targeted source-language phenomenon as the primary translation challenge, without being obscured by additional difficulties in other parts of the sentence, insofar as this is possible.

The current scoring approach does not differentiate sources by their translation difficulty. In future work, we plan to introduce a weighting scheme informed by the entropy of submitted translation solutions.

Acknowledgments

We would like to thank Vladimir Kropivnitskiy for his contribution to the manual annotation of the test suite this year. We would also like to thank Vivien Macketanz, Sergei Bagdasarov, Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai, and He Wang for their prior contributions to the creation of the test suite.

References

Lars Ahrenberg. 2017. Comparing machine translation and human translation: A case study. In *Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, pages 21–28, Varna, Bulgaria.

Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2022. A Test Suite for the Evaluation of Portuguese-English Machine Translation. In *Computational Processing of the Portuguese Language*, pages 15–25, Cham. Springer International Publishing.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*. Association for Computational Linguistics.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Singapore. Association for Computational Linguistics.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2024. Machine translation metrics are better in evaluating linguistic errors on LLMs than on encoder–decoder systems. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

Hanna Bittner. 2020. *Evaluating the Evaluator: A Novel Perspective on Translation Quality Assessment*. Routledge, New York and London.

Michael Carl and Moritz Jonas Schaeffer. 2017. Why translation is difficult: A corpus-based study of non-literality in post-editing and from-scratch translation. *Hermes (Denmark)*, 56:43–57.

Gys-Walt Van Egdom, Heidi Verplaetse, Iris Schrijver, Hendrik J. Kockaert, Winibert Segers, Jasper Pauwels, Bert Wylin, and Henri Bloemen. 2019. How to Put the Translation Test to the Test? On Preselected Items Evaluation and Perturbation. In Elsa Huertas-Barros, Sonia Vandepitte, and Emilia Iglesias-Fernández, editors, *Quality Assurance and Assessment Practices in Translation and Interpreting*, pages 26–56. IGI Global, Hershey, PA, USA.

Mara Finkelstein, Geza Kovacs, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Markus Freitag, and David Vilar. 2025. Google Translate's Research Submission to WMT2025. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Nikolay Karpachev, Ekaterina Enikeeva, Dmitry Popov, Arsenii Bulgakov, Daniil Panteleev, Dmitrii Ulianov, Artem Kryukov, and Artem Mekhraliev. 2025. Yandex Submission to the WMT25 General Machine Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, Nithya Govindarajan, Wei-Yin Ko, Julia Kreutzer, Kelly Marchisio, Ahmet Üstün, Sebastian Vincent, and Ivan Zhang. 2025a. Command-A-Translate: Raising the Bar of Machine Translation with Difficulty Filtering. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025b. Preliminary Ranking of WMT25 General Machine Translation Systems.

Maria Kunilovskaya, Tatyana Ilyushchenya, Natalia Morgoun, and Ruslan Mitkov. 2023. Source language difficulties in learner translation: Evidence from an error-annotated corpus. *Target*, 35(1):34–62.

Maria Kunilovskaya, Iuliia Zaitova, Wei Xue, Irina Stenger, and Tania Avgustinova. 2025. Predictability of microsyntactic units across slavic languages: A translation-based study. In *The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*, pages 313–322.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art Machine Translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation. (WMT21)*, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371, Miami, Florida, USA. Association for Computational Linguistics.

Christiane Nord. 1997. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. St. Jerome, Manchester, UK.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text. A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter, Berlin.

Hao Wang. 2025. Wenyiil's Submissions to the WMT 2025 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Yuxiang Wei. 2022. Entropy as a measurement of cognitive load in translation. In *AMTA 2022 - 15th Conference of the Association for Machine Translation in the Americas, Proceedings - Workshop on Empirical Translation Process Research*, volume 1, pages 75–86.

Di Wu, Yan Meng, Maya Konstantinovna Nachesa, Seth Aycock, and Christof Monz. 2025. UvA-MT's Participation in the WMT25 General Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Linlong Xu. 2025. Algharb at WMT25 Translation Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-MT at WMT25 General Machine Translation Shared Task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

categ	count	Wenyi	Algha	Yande	Gemin	Claud	DeepS	GPT41	Shy	Comma	UvAMT	avg
Ambiguity	10	100.0	100.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	97.0
Coordination & ellipsis	27	85.2	85.2	92.6	88.9	85.2	81.5	85.2	88.9	96.3	88.9	87.8
False friends	8	100.0	100.0	100.0	87.5	100.0	87.5	100.0	75.0	75.0	100.0	92.5
Function word	16	93.8	93.8	87.5	93.8	100.0	100.0	93.8	93.8	87.5	93.8	93.8
LDD & interrogatives	27	96.3	96.3	88.9	92.6	88.9	85.2	81.5	92.6	85.2	85.2	89.3
Lexical Morphology	16	87.5	87.5	93.8	93.8	87.5	87.5	87.5	87.5	75.0	93.8	88.1
MWE	55	**85.5**	**85.5**	**92.7**	80.0	78.2	78.2	**83.6**	**83.6**	69.1	67.3	80.4
Named entity & terminology	47	**91.5**	**91.5**	**97.9**	**93.6**	**93.6**	**91.5**	**95.7**	**89.4**	**89.4**	85.1	91.9
Non-verbal agreement	39	**100.0**	**100.0**	89.7	**97.4**	92.3	**94.9**	92.3	92.3	**94.9**	89.7	94.4
Subordination	49	**98.0**	**98.0**	**95.9**	**100.0**	93.9	93.9	93.9	**95.9**	87.8	85.7	94.3
Verb semantics	7	**100.0**	**100.0**	**100.0**	**85.7**	**71.4**	**85.7**	**71.4**	**71.4**	**85.7**	42.9	81.4
Verb tense/aspect/mood	94	**94.7**	**94.7**	**94.7**	**94.7**	**89.4**	**92.6**	**90.4**	**91.5**	**93.6**	87.2	92.3
Verb valency	70	91.4	91.4	90.0	87.1	90.0	87.1	87.1	87.1	81.4	81.4	87.4
micro-average	465	**93.1**	**93.1**	**93.1**	**91.8**	89.5	89.2	89.5	89.7	86.5	83.7	89.9
macro-average	465	**94.1**	**94.1**	**93.4**	**91.9**	90.0	89.7	89.4	88.4	86.2	83.1	90.0
our rank		1	1	1	1	5	5	5	5	9	10	
WMT25 human rank		3	5	8	1	3	6	3	2	6	10	

phenomenon	count	Wenyi	Algha	Yande	Gemin	Claud	DeepS	GPT41	Shy	Comma	UvAMT	avg
Ambiguity	10	100.0	100.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	97.0
Lexical ambiguity	10	100.0	100.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	97.0
Coordination & ellipsis	27	85.2	85.2	92.6	88.9	85.2	81.5	85.2	88.9	96.3	88.9	87.8
Gapping	5	80.0	80.0	100.0	100.0	80.0	80.0	60.0	60.0	80.0	60.0	78.0
Pseudogapping	7	71.4	71.4	85.7	100.0	85.7	71.4	100.0	100.0	100.0	100.0	88.6
Right node raising	5	80.0	80.0	100.0	80.0	80.0	80.0	80.0	80.0	100.0	80.0	84.0
Sluicing	2	100.0	100.0	100.0	50.0	100.0	50.0	50.0	100.0	100.0	100.0	85.0
Stripping	6	100.0	100.0	83.3	83.3	83.3	100.0	100.0	100.0	100.0	100.0	95.0
VP-ellipsis	2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
False friends	8	100.0	100.0	100.0	87.5	100.0	87.5	100.0	75.0	75.0	100.0	92.5
Function word	16	93.8	93.8	87.5	93.8	100.0	100.0	93.8	93.8	87.5	93.8	93.8
Focus particle	4	75.0	75.0	50.0	75.0	100.0	100.0	75.0	75.0	50.0	100.0	77.5
Question tag	12	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	91.7	99.2
LDD & interrogatives	27	96.3	96.3	88.9	92.6	88.9	85.2	81.5	92.6	85.2	85.2	89.3
Inversion	8	87.5	87.5	87.5	87.5	87.5	87.5	75.0	100.0	75.0	87.5	86.3
Multiple connectors	3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	66.7	100.0	100.0	96.7
Pied-piping	7	100.0	100.0	85.7	100.0	85.7	85.7	85.7	100.0	85.7	100.0	92.9
Preposition stranding	3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Topicalization	3	100.0	100.0	66.7	66.7	100.0	66.7	66.7	100.0	100.0	66.7	83.3
Wh-movement	3	100.0	100.0	100.0	100.0	66.7	66.7	66.7	66.7	66.7	33.3	76.7
Lexical Morphology	16	87.5	87.5	93.8	93.8	87.5	87.5	87.5	87.5	75.0	93.8	88.1
Functional shift	8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Noun formation (er)	8	75.0	75.0	87.5	87.5	75.0	75.0	75.0	75.0	50.0	87.5	76.3
MWE	55	**85.5**	**85.5**	**92.7**	80.0	78.2	78.2	**83.6**	**83.6**	69.1	67.3	80.4
Collocation	9	88.9	88.9	88.9	88.9	77.8	77.8	77.8	88.9	66.7	66.7	81.1
Compound	6	66.7	66.7	66.7	33.3	50.0	66.7	66.7	66.7	66.7	50.0	60.0
Idiom	13	**92.3**	**92.3**	**100.0**	**84.6**	**84.6**	76.9	**100.0**	69.2	61.5	76.9	83.8
Nominal MWE	8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	50.0	93.8
Prepositional MWE	7	100.0	100.0	100.0	100.0	71.4	100.0	85.7	100.0	85.7	71.4	91.4
Verbal MWE	12	66.7	66.7	91.7	66.7	75.0	58.3	66.7	83.3	58.3	75.0	70.8
Named entity & terminology	47	**91.5**	**91.5**	**97.9**	**93.6**	**93.6**	**91.5**	**95.7**	**89.4**	**89.4**	85.1	91.9
Date	17	88.2	88.2	100.0	94.1	94.1	88.2	100.0	94.1	100.0	94.1	94.1
Domainspecific Term	2	50.0	50.0	100.0	50.0	50.0	50.0	50.0	50.0	100.0	50.0	60.0
Measuring Unit	9	88.9	88.9	100.0	100.0	100.0	100.0	100.0	88.9	88.9	100.0	95.6
Onomatopeia	4	100.0	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	95.0
Proper Name & Location	15	**100.0**	**100.0**	**100.0**	**93.3**	**93.3**	**93.3**	**93.3**	**86.7**	73.3	73.3	90.7
Non-verbal agreement	39	**100.0**	**100.0**	89.7	**97.4**	92.3	**94.9**	92.3	92.3	**94.9**	89.7	94.4
Coreference	14	**100.0**	**100.0**	**85.7**	**92.9**	**85.7**	**92.9**	**85.7**	**92.9**	**92.9**	78.6	90.7
Genitive	10	100.0	100.0	100.0	100.0	90.0	90.0	90.0	90.0	90.0	90.0	94.0
Personal Pronoun Coreference	4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Substitution	11	100.0	100.0	81.8	100.0	100.0	100.0	100.0	90.9	100.0	100.0	97.3
Subordination	49	**98.0**	**98.0**	**95.9**	**100.0**	93.9	93.9	93.9	**95.9**	87.8	85.7	94.3
Adverbial clause	1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Cleft sentence	3	100.0	100.0	66.7	100.0	66.7	100.0	66.7	100.0	66.7	33.3	80.0
Complex object	9	88.9	88.9	100.0	100.0	88.9	77.8	88.9	88.9	88.9	88.9	90.0
Contact clause	3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Infinitive clause	12	**100.0**	**100.0**	**91.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**91.7**	95.8
Object clause	3	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Participle clause	12	100.0	100.0	100.0	100.0	91.7	100.0	91.7	91.7	91.7	83.3	95.0
Subject clause	6	100.0	100.0	100.0	100.0	100.0	83.3	100.0	100.0	100.0	83.3	96.7
Verb semantics	7	**100.0**	**100.0**	**100.0**	**85.7**	**71.4**	**85.7**	**71.4**	**71.4**	**85.7**	42.9	81.4
Verb tense/aspect/mood	94	**94.7**	**94.7**	**94.7**	**94.7**	**89.4**	**92.6**	**90.4**	**91.5**	**93.6**	87.2	92.3
Conditional	12	100.0	100.0	100.0	100.0	91.7	100.0	100.0	100.0	100.0	91.7	98.3
Ditransitive	22	**100.0**	**100.0**	86.4	**100.0**	**95.5**	**95.5**	**100.0**	86.4	**100.0**	86.4	95.0
Gerund	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Imperative	16	**87.5**	**87.5**	**100.0**	**87.5**	81.3	**87.5**	81.3	**87.5**	**93.8**	**87.5**	88.1
Intransitive	22	90.9	90.9	90.9	90.9	86.4	90.9	86.4	90.9	95.5	86.4	90.0
Reflexive	11	90.9	90.9	100.0	90.9	90.9	90.9	90.9	90.9	81.8	90.9	90.9
Transitive	6	100.0	100.0	100.0	100.0	83.3	83.3	66.7	100.0	66.7	66.7	86.7
Verb valency	70	91.4	91.4	90.0	87.1	90.0	87.1	87.1	87.1	81.4	81.4	87.4
Case government	20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Catenative verb	13	**92.3**	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	**100.0**	**92.3**	76.9	**84.6**	91.5

phenomenon	count	Wenyi	Algha	Yande	Gemin	Claud	DeepS	GPT41	Shy	Comma	UvAMT	avg
Mediopassive voice	6	100.0	100.0	83.3	83.3	100.0	100.0	100.0	83.3	66.7	66.7	88.3
Passive voice	12	100.0	100.0	100.0	100.0	100.0	91.7	91.7	100.0	100.0	100.0	98.3
Resultative	8	**100.0**	**100.0**	**87.5**	62.5	**87.5**	**87.5**	**75.0**	**87.5**	**75.0**	**75.0**	83.8
Semantic roles	11	54.5	54.5	63.6	63.6	45.5	45.5	45.5	45.5	45.5	36.4	50.0
micro-average	465	**93.1**	**93.1**	**93.1**	**91.8**	89.5	89.2	89.5	89.7	86.5	83.7	89.9
phen. macro-average	465	**93.0**	**93.0**	**92.5**	**90.5**	89.0	88.3	87.4	89.4	86.6	82.4	89.2
categ. macro-average	465	**94.1**	**94.1**	**93.4**	**91.9**	90.0	89.7	89.4	88.4	86.2	83.1	90.0

Tagged Span Annotation for Detecting Translation Errors in Reasoning LLMs

Taemin Yeom[1,2], Yonghyun Ryu[2], Yoonjung Choi[2†], JinYeong Bak[3†]

[1]Department of Digital Media and Communications Engineering, Sungkyunkwan University,
[2]Samsung Research,
[3]Department of Computer Science and Engineering, Sungkyunkwan University
taemin.yeom@g.skku.edu
{yonghyun.ryu, yj0807.choi}@samsung.com
jy.bak@skku.edu

Abstract

We present the submission of the AIP team to the WMT 2025 Unified MT Evaluation Shared Task, focusing on the span-level error detection subtask. Our system emphasizes response-format design to better harness the capabilities of OpenAI's o3, the state-of-the-art reasoning LLM. To this end, we introduce Tagged Span Annotation (TSA), an annotation scheme designed to more accurately extract span-level information from the LLM. On our refined version of WMT24 ESA dataset, our reference-free method achieves an F1 score of approximately 27 for character-level label prediction, outperforming the reference-based XCOMET-XXL at approximately 17.[1]

1 Introduction

With the recent widespread use of the LLM-as-a-judge approach, research on human-like translation quality evaluation using large language models (LLMs)—such as GEMBA-DA (Kocmi and Federmann, 2023b), MQM (Kocmi and Federmann, 2023a), ESA (Zouhar et al., 2025), EAPrompt (Lu et al., 2024), and AutoMQM (Fernandes et al., 2023)—has grown rapidly. Fine-grained error detection enables explainable translation evaluation and informs the design of post-editing systems, making it widely applicable across diverse systems that leverage machine translation. Nevertheless, research on fine-grained translation error detection remains scarce; only a handful of studies exist despite growing interest. Moreover, the LLM-based studies mentioned above focused less on fine-grained error detection itself and more on using it to compute final machine translation (MT) evaluation scores. Additionally, most prior work (Kocmi and Federmann, 2023b,a; Fernandes et al., 2023) has been limited to non-reasoning LLMs. With the recent emergence of reasoning models that have achieved state-of-the-art performance across a wide range of tasks, there is, therefore, a need to investigate their use for fine-grained error detection as well.

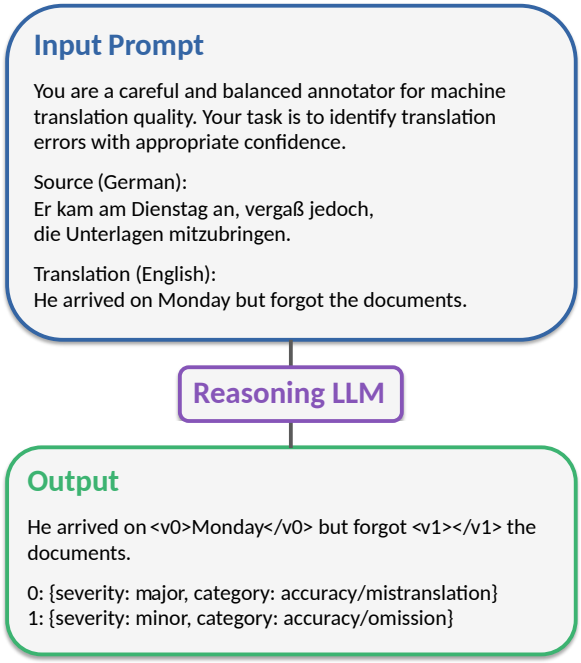


Figure 1: The overview of **Tagged Span Annotation (TSA)** system. The reference translation is *He arrived on Tuesday, but forgot to bring the documents.* The LLM tags each error span in the hypothesis with inline tags <vN> and returns a severity–category pair for every tag.

In this paper, we propose a translation error span detection system that leverages OpenAI o3 (OpenAI, 2025a), the state-of-the-art reasoning large language model, for the WMT'25 MT evaluation span-level error detection task. Our system adopts

[†]Corresponding authors
[1]Code and dataset repo: https://github.com/TaeminYeom/Tagged_Span_Annotation

Proceedings of the Tenth Conference on Machine Translation, pages 878–886
November 8-9, 2025

Lang	TSA (Ours)			XCOMET-XXL-QE			XCOMET-XXL			GEMBA			GEMBA (fixed)		
	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.
en-cs	**25.24**	**23.32**	27.49	15.07	9.73	**33.42**	15.90	10.44	33.33	4.25	12.13	2.57	18.26	14.24	25.44
en-ja	**32.15**	**21.91**	**60.35**	7.84	4.69	23.98	8.63	5.24	24.46	1.42	1.70	1.21	5.73	3.66	13.20
en-zh	**17.86**	**13.07**	28.21	7.32	4.07	**36.03**	7.58	4.25	35.24	4.11	5.38	3.33	9.72	6.35	20.65
en-is	32.09	**47.66**	24.19	33.77	30.74	37.46	**35.48**	33.27	**38.01**	1.19	26.43	0.61	30.95	38.68	25.79
en-uk	**31.10**	**29.52**	**32.85**	15.83	11.52	25.28	17.01	12.72	25.67	1.48	7.69	0.82	14.92	13.41	16.82
en-ru	**23.87**	**28.54**	20.51	18.05	16.12	20.50	19.18	16.85	**22.26**	3.79	18.85	2.11	19.39	19.59	19.19
Avg.	**27.05**	**27.34**	**32.27**	16.31	12.81	29.45	17.30	13.79	29.83	2.71	12.03	1.78	16.50	15.99	20.18

Table 1: Span-level detection performance of TSA method compared with baseline systems on WMT24 ESA dataset six language pairs. TSA consistently achieves the highest average F1, Precision, and Recall in most languages.

structured output to ensure reliable parsing of responses and represents error locations using tagged-span annotations. On our refined version of the WMT24 ESA dataset, our reference-free system achieved an F1 score of 27.05, outperforming the reference-based XCOMET-XXL at 17.30.

2 Method

2.1 Background

Transformer-encoder approaches such as XCOMET (Guerreiro et al., 2024) typically cast span-level error detection as a token-classification task on the translation, assigning a severity label—*no-error*, *minor*, or *major*—to every token.

In generative-response settings, error spans can be expressed in three canonical ways:

(i) by returning the erroneous text itself

(ii) by providing its start- and end-character indices

(iii) by inserting inline tags directly into the translation

Method (i) is the format adopted by both GEMBA-MQM and GEMBA-ESA. In GEMBA-ESA, the returned span string is used only as a hint, and in MQM the sentence-level score is computed solely from the number of spans and their severity labels. Hence, for *sentence-level scoring* this representation is sufficient. For a *span-detection* task, however, it is inadequate: the same substring can occur multiple times in a sentence, so the actual error location remains ambiguous.

Method (ii)—returning start and end character indices looks attractive for its simplicity and precision. In practice, however, generative LLMs struggle with producing *exact* numerical values: they operate on sub-word (BPE) tokens rather than raw characters, so mapping token boundaries back to UTF-8 offsets is non-trivial (Zhang and He, 2024)

Method (iii) follows the inline-tag scheme used in the publicly released *WMT MQM Human-Evaluation* dataset (Freitag et al., 2021). In that corpus, each record contains exactly one error span; when a translation exhibits multiple errors, it is split into multiple records, each highlighting a single span.

2.2 Tagged Span Annotation (TSA)

In our system we adopt **method (iii)** and extend it with *numbered inline tags*: each error span is wrapped in a unique <v$_k$>–</v$_k$> pair, allowing multiple errors to be annotated simultaneously within the same sentence.

When a source segment is omitted in the hypothesis, we mark the omission by inserting a zero-length tag pair <v$_k$></v$_k$> at the exact insertion point.

To elicit the desired response format, we composed a detailed system prompt, supplied few-shot examples that explicitly illustrate the target schema, and leveraged the model's *structured-output* interface to guarantee machine-parseable answers. (see Appendix A; OpenAI, 2023a).

We evaluate three OpenAI models—**GPT-4.1** (OpenAI, 2025b), **o3** (OpenAI, 2025a), and **o4-mini** (OpenAI, 2025a). Decoding parameters follow the API defaults, except that GPT-4.1 uses a lower temperature of 0.2 to reduce variance.

3 Experiments

3.1 Datasets

We use the our refined version of **WMT24** ESA Annotations datasets[2] for six language pairs: **en→{cs,**

[2] https://github.com/wmt-conference/wmt24-news-systems

is, ja, ru, uk, zh}. Each pair contains approximately **6–8k** segments, for a total of about **39,685** segments. We preserve the human **ESA gold** annotations as the reference for span-level evaluation. [3]

3.2 Metrics

We follow the official WMT25 Task 2 ESA scorer[4]. Span detection is evaluated by **character-level overlap**, not span matching: for each character position, the counts of *major/minor* errors in gold and prediction are compared, and the true-positive score is calculated as the sum of the minimum counts at each character position. Partial overlaps that match in location but not in severity receive **partial credit** (0.5). Omissions marked as `start="missing"` are ignored by the scorer. MQM categories are not used for scoring. We report precision, recall, and F1 as micro-averages over segments within each language, and then macro-averages across languages.

3.3 Baselines

We compare our methods against two established baselines: **XCOMET** (Guerreiro et al., 2024) and **GEMBA** (Kocmi and Federmann, 2023a). Both XCOMET and GEMBA are also used as official baselines in the WMT25 Shared Task. All scores are character-level and macro-averaged across the six MQM language pairs; within each language we compute micro averages over segments. For GEMBA, we adhere to the original prompt and settings with one modification: the few-shot prompt is revised to require outputs in a strict JSON format to facilitate reliable parsing. In preliminary experiments, the unmodified prompt frequently elicited explanatory text in addition to the JSON output, resulting in parsing failures and degraded performance. To mitigate this issue, we explicitly instructed the model to omit any explanatory content, which yielded improved performance. We denote this variant as GEMBA (fixed).

[3]The script to refine the WMT24 ESA has been merged into the WMT25 official repository: https://github.com/wmt-conference/wmt25-mteval/blob/3332614/scripts/devset/create_tsv_from_wmt24_esa.sh

[4]https://github.com/wmt-conference/wmt25-mteval/blob/3332614/scripts/scoring/task2/scoring_esa.py

Method	F1	Prec.	Rec.
GEMBA	19.70	20.70	21.54
Direct-index	22.57	27.32	23.50
TSA (no precision emphasis)	25.43	21.99	**38.19**
TSA	**27.05**	**27.34**	32.27

Table 2: Ablation on output format and precision-emphasis prompting on o3 model. GEMBA, Direct-index, and TSA use an **identical prompt**, differing only in the directive that specifies the output format. *TSA (no precision emphasis)* employs the same TSA format but removes the precision-oriented instruction, isolating its contribution.

4 Results

Main results Table 1 presents the primary results. Our best system combines the **o3** reasoning model with the *Tagged Span Annotation* output design and achieves the highest F1 score. Under the same evaluation setting, it surpasses the *reference*-based **XCOMET-XXL** by **+9.65** F1 and the **GEMBA (fixed)** baseline by **+10.55** F1.

4.1 Ablation Study

We perform extensive experiments to identify the factors that influence response quality, and we report our findings here.

4.1.1 Span Annotation

We compare three output formats (i) **TSA**, (ii) a **GEMBA** format that returns the error-span text itself and (iii) a **Direct-index** format that outputs the character indices of each span using the **o3** model.

Except for the directive specifying the output format, every part of the prompt is kept identical, allowing us to isolate performance differences attributable solely to the span-annotation scheme. As shown in Table 2, the performance gap between the GEMBA and Direct-index baselines and our TSA method confirms this effect.

4.1.2 Precision Emphasis

In the TSA setting, we observe a pronounced precision–recall imbalance: recall was consistently much higher than precision in most language pairs.

Because the F1 score is the harmonic mean of precision and recall, it is dominated by the lower of the two components; consequently, it attains higher values only when the two metrics are balanced. To mitigate the observed imbalance, we explicitly

Method	GPT-4.1			o3			o4-mini		
	F1	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**
GEMBA	**21.16**	**20.84**	21.16	19.70	20.70	21.54	17.42	15.28	23.50
Direct-index	13.46	14.92	14.95	22.57	27.32	23.50	18.13	19.30	20.76
TSA	20.03	18.45	**26.82**	**27.05**	**27.34**	**32.27**	**21.26**	**19.95**	**26.86**

Table 3: Span-level detection Performance by model and output format. Tagged Span Annotation (TSA) shows no noticeable performance gains over GEMBA or Direct-index on the non-reasoning baseline `GPT-4.1`, but it delivers substantial improvements on the reasoning models `o3` and `o4-mini`, highlighting the pronounced benefit of span markup for models with stronger reasoning capability.

reinforced the requirement of promoting higher precision.

We instruct the model to assign an error label only when it is *confident* the translation is incorrect. Moreover, because the model tends to select spans wider than the actual error, the prompt explicitly requires it to tag only the *minimal* substring that covers the core error.

For comparison, Table 2 also reports a TSA variant in which the prompt *did not* include the precision-oriented instructions.

4.1.3 Reasoning Impact

As shown in Table 3, the gap between TSA and the GEMBA format is most pronounced for **reasoning models**. We interpret that this stems from the two-stage decision process enforced by TSA, which closely resembles the human ESA scoring protocol: human annotators **first** indicate the mark spans and severity of errors, and then score the translation quality. ESA has been reported to achieve inter-annotator agreement comparable to full MQM annotation (Kocmi et al., 2024). Similarly, TSA requires the LLM to mark error spans in the `annotated_translation` field before predicting their severity and type, thus externalising an intermediate reasoning step.

The non-reasoning baseline GPT-4.1 shows no performance improvement from TSA over GEMBA, whereas the reasoning models `o3` and `o4-mini` exhibit substantial gains. This suggests that models with stronger reasoning ability are better able to exploit the intermediate span markup to conduct a more fine-grained error analysis.

5 Related Work

5.1 Encoder-Based Evaluation

COMET (Rei et al., 2020) is the first to attach a *regression head* to a pretrained cross-lingual encoder (e.g., XLM-R) in order to predict Direct Assessment (DA) scores, achieving higher human correlation than traditional metrics such as BLEU (Papineni et al., 2002) and TER (Snover et al., 2006). **COMET-MQM** (Rei et al., 2021) added an *auxiliary token-classification task* that predicts whether each token falls inside an error span based on MQM annotations.

XCOMET (Guerreiro et al., 2024) unifies and generalizes this line of work by (i) retaining the DA regression head and (ii) replacing COMET-MQM's token classifier with a *span head* that emits start–end indices. Yet these index-based spans lack error *types/categories*, limiting diagnostic power and compatibility with MQM/ESA labels, and the primary training objective remains sentence-level scoring, leaving boundary accuracy secondary. We treat XCOMET-XXL as the strongest encoder-based baseline and, inspired by its index span annotation, test an Direct-index format variant in our LLM-as-judge pipeline (§4.1.1); however, vulnerability to character-index drift ultimately leads us to adopt *Tagged Span Annotation (TSA)* for our final design.

5.2 LLM-as-a-judge

GEMBA (Kocmi and Federmann, 2023a) frames MT evaluation as an LLM-as-judge task: a prompted GPT-4 produces Direct Assessment (DA) scores or MQM/ESA error spans and severities from the source and translation (optionally a reference). However, GEMBA's few-shot, free-form outputs can yield duplicated instructions across exemplars, brittle parsing. To improve robustness, we convert the output to a JSON format and adopted structured-output prompting.

MQM-APE (Lu et al., 2025) augments GEMBA with a downstream *automatic post-editing* (APE) pass: after the initial GPT-4 annotation, a second LLM rewrites the MT segment with the pro-

posed fixes and discards spans whose removal does not change the edited meaning. This post-filtering raises span precision with little effect on recall. Because MQM-APE operates after span generation, it is orthogonal to our TSA, JSON-validated spans (with minimal fixes) could be plugged into MQM-APE without modifying its algorithm. We leave a full integration and evaluation to future work.

6 Limitations

This study has the following limitations. First, the experiments were conducted primarily on proprietary large language models (LLMs) such as `gpt-4.1`, `o3` and `o4-mini`, without providing comparable results for large-scale open-source models. This limits the generalizability of the proposed approach to other model families. Second, although per-language results were reported, we did not conduct in-depth analyses of performance variations or error patterns across specific language pairs. In particular, the causes of performance differences for low-resource languages remain underexplored. Third, the monetary cost of using proprietary LLMs can hinder large-scale adoption. Using `tiktoken` (OpenAI, 2023b) to count tokens, our evaluation on six WMT24 ESA language pairs (39,684 segments, 56.82M input tokens, 3.26M output tokens) is estimated to have cost about $139 in total ($3.5 per 1,000 segments) under the current `o3` pricing ($2/M input, $8/M output as of September 2025), which may hinder large-scale adoption despite the accuracy gains.

7 Conclusion and Future Work

We introduced **Tagged Span Annotation** (TSA), a structured-output framework that pairs numbered inline tags with a JSON schema to enable reliable, fine-grained translation-error detection. Across six WMT language pairs, TSA achieved the highest span-level F1 scores, surpassing the strongest GEMBA variant by +10.55 and XCOMET-XXL by +9.65 absolute points on average.

Reasoning vs. non-reasoning models. We observed performance differences between reasoning and non-reasoning LLMs. As future work, we will systematically study how prompting style and intermediate reasoning affect span detection by comparing: (i) direct answers vs. chain-of-thought (CoT; Wei et al. 2022) prompting, (ii) single-pass decoding vs. self-consistency with multiple CoT paths, and (iii) unguided CoT vs. CoT *path selection* aided by a span-level verifier. We will evaluate not only span F1 (overall and by severity/type), but also invalid-output rate, span-length bias, latency, and cost to characterize accuracy–efficiency trade-offs.

Open-source models trained on evaluation data. To improve reproducibility and broaden applicability, we will fine-tune strong open-source LLMs on human-annotated MQM/ESA data (and carefully curated synthetic data), comparing SFT vs. parameter-efficient adapters (e.g., LoRA; Hu et al. 2021) under multilingual vs. per-language regimes. We will analyze cross-lingual transfer (high- to low-resource), domain shift, and calibration quality, and report effect sizes alongside standard correlations. Where licensing permits, we plan to release training code, prompts, and evaluation harnesses to facilitate future benchmarking.

Coupling with post-editing (MQM-APE). We further plan to couple our span detector with post-editing systems that explicitly target MQM categories (e.g., MQM-aware APE). Concretely, we will explore (i) using predicted spans and severities to guide edit proposals, (ii) verifier- or reward-based reranking of edits, and (iii) joint training where an APE loss encourages span-consistent corrections. We will report pre-/post-edit score deltas, human correlation, and downstream adequacy/fluency gains to quantify the benefit of integrating detection with correction.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful questions and comments. JinYeong Bak was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-RS-2025-00523385).

References

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of

human evaluation for machine translation. *Preprint*, arXiv:2104.14478.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *Preprint*, arXiv:2406.11580. ArXiv preprint.

Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5570–5587, Abu Dhabi, UAE. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2023a. Function calling and other api updates. https://openai.com/index/function-calling-and-other-api-updates/. Accessed 2025-09-26.

OpenAI. 2023b. tiktoken. https://github.com/openai/tiktoken. Accessed 2025-09-26.

OpenAI. 2025a. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed 2025-09-26.

OpenAI. 2025b. System card: Gpt-4.1. https://openai.com/index/gpt-4-1/. Accessed 2025-09-26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yidan Zhang and Zhenan He. 2024. Large language models can not perform well in understanding and manipulating natural language at both character and word levels? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11826–11842, Miami, Florida, USA. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

A Prompt

A.1 System Prompt

We used the following system prompt. It described the essential rules for marking error spans and placed emphasis on addressing the issue of precision being significantly lower than recall.

```
You are a careful and balanced annotator for machine translation quality.
Your task is to identify translation errors with appropriate confidence.

## EVALUATION GUIDELINES:
- Be thorough but precise: Only mark errors when you are confident they are
↪  incorrect
- Consider context and domain: Some variations may be acceptable depending on
↪  context
- Distinguish between errors and acceptable alternatives: Multiple valid
↪  translations may exist
- Focus on clear, objective errors rather than subjective preferences
- Verify each potential error against the source text before marking
- When in doubt, err on the side of not marking an error

## Error Categories:
- Accuracy: addition, mistranslation, omission, untranslated text
- Fluency: character encoding, grammar, inconsistency, punctuation, register,
↪  spelling
- Style: awkward phrasing
- Terminology: inappropriate for context, inconsistent use
- Other: non-translation, other issues

## Severity Classification:
- Major: Errors that impact meaning or usability but do not render the text unusable
- Minor: Errors that do not impact meaning or usability

## CRITICAL OUTPUT REQUIREMENTS:
- Mark errors only when you have clear evidence they are incorrect
- Consider whether alternative translations could be equally valid
- Apply strict standards: better to miss a minor error than create a false positive
- Wrap error spans in the translation with tags: <v0>, <v1>, <v2>, etc.
- Use tag numbers in sequential order starting from <v0>. Do not skip numbers or use
↪  them out of order
- For omissions, place empty tags <vN></vN> where the missing text should have been
- NO comments, explanations, or additional text
- Mark only the minimal substring that contains the clear error; do not include
↪  extra context
```

A.2 Few-shot Examples

We revised the GEMBA-MQM (Kocmi and Federmann, 2023a) few-shot examples to align with the TSA format. In addition, we added one example without any error spans.

```
Source (English):
```

```
I do apologise about this, we must gain permission from the account holder to
↪  discuss an order with another person, I apologise if this was done previously,
↪  however, I would not be able to discuss this with yourself without the account
↪  holders permission.

Translation (German):
Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung
↪  mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor
↪  geschehen wäre, aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der
↪  Lage, dies mit dir involvement.

Answer:
{
"annotated_translation": "Ich entschuldige mich dafür, wir müssen die
↪  Erlaubnis<v0></v0> einholen, um eine Bestellung mit einer anderen Person zu
↪  besprechen. Ich entschuldige mich, falls dies zuvor geschehen <v1>wäre</v1>,
↪  aber ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit
↪  <v2>dir</v2> <v3>involvement</v3>.",
"errors": [
    {"severity": "Major", "category": "accuracy/omission"},
    {"severity": "Minor", "category": "fluency/grammar"},
    {"severity": "Minor", "category": "fluency/register"},
    {"severity": "Major", "category": "accuracy/mistranslation"}
]
}

---

Source (English):
Talks have resumed in Vienna to try to revive the nuclear pact, with both sides
↪  trying to gauge the prospects of success after the latest exchanges in the
↪  stop-start negotiations.

Translation (Czech):
Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemž obě
↪  partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.

Answer:
{
"annotated_translation": "Ve Vídni se <v0>ve Vídni</v0> obnovily rozhovory o
↪  oživení jaderného paktu, přičemž obě <v1>partaje</v1> se snaží posoudit
↪  vyhlídky na úspěch po posledních výměnách v<v2></v2> jednáních.",
"errors": [
    {"severity": "Major", "category": "accuracy/addition"},
    {"severity": "Minor", "category": "terminology/inappropriate for context"},
    {"severity": "Major", "category": "accuracy/omission"}
]
}

---

Source (Chinese):
```

```
大点木家居道提供高居然之家地址，等最新商信息，修公司，就上大点

Translation (English):
Urumqi Home Furnishing Store Channel provides you with the latest business
↪  information such as the address, telephone number, business hours, etc., of
↪  high-speed rail, and find a decoration company, and go to the reviews.

Answer:
{
"annotated_translation": "Urumqi Home Furnishing Store Channel provides you with
↪  the latest business information such as the address, telephone number, business
↪  hours, <v0>etc.,</v0> <v1>of high-speed rail,</v1> and find a decoration
↪  company, and <v2>go to the reviews</v2>.",
"errors": [
    {"severity": "Minor", "category": "style/awkward"},
    {"severity": "Major", "category": "accuracy/addition"},
    {"severity": "Major", "category": "accuracy/mistranslation"}
]
}

---

Source (English):
According to the terms outlined in the agreement, the supplier shall deliver all
↪  components no later than thirty days after receiving the initial purchase order,
↪  and any delays must be communicated in writing at least five business days in
↪  advance.

Translation (Spanish):
De acuerdo con los términos establecidos en el acuerdo, el proveedor deberá
↪  entregar todos los componentes a más tardar treinta días después de recibir la
↪  orden de compra inicial, y cualquier retraso deberá comunicarse por escrito con
↪  al menos cinco días hábiles de antelación.

Answer:
{
"annotated_translation": "De acuerdo con los términos establecidos en el acuerdo,
↪  el proveedor deberá entregar todos los componentes a más tardar treinta días
↪  después de recibir la orden de compra inicial, y cualquier retraso deberá
↪  comunicarse por escrito con al menos cinco días hábiles de antelación.",
"errors": [
]
}
```

COMET-poly: Machine Translation Metric Grounded in Other Candidates

Maike Züfle[1]★ **Vilém Zouhar**[2]★ **Tu Anh Dinh**[1]★ **Felipe Maia Polo**[3]
Jan Niehues[1] **Mrinmaya Sachan**[2]

[1]Karlsruhe Institute of Technology [2]ETH Zurich [3]University of Michigan
{maike.zuefle,tu.dinh}@kit.edu vzouhar.ethz.ch

Abstract

Automated metrics for machine translation attempt to replicate human judgment. Unlike humans, who often assess a translation in the context of multiple alternatives, these metrics typically consider only the source sentence and a single translation. This discrepancy in the evaluation setup may negatively impact the performance of automated metrics. We propose two automated metrics that incorporate additional information beyond the single translation. $\text{COMET}_{\text{poly-cand}}$ uses alternative translations of the same source sentence to compare and contrast with the translation at hand, thereby providing a more informed assessment of its quality. $\text{COMET}_{\text{poly-ic}}$, inspired by retrieval-based in-context learning, takes in translations of similar source texts along with their human-labeled quality scores to guide the evaluation. We find that including a single additional translation in $\text{COMET}_{\text{poly-cand}}$ improves the segment-level metric performance (0.079→0.118 τ_b), with further gains when more translations are added. Incorporating retrieved examples in $\text{COMET}_{\text{poly-ic}}$ yields similar improvements (0.079→0.116 τ_b). We release our models publicly.[1]

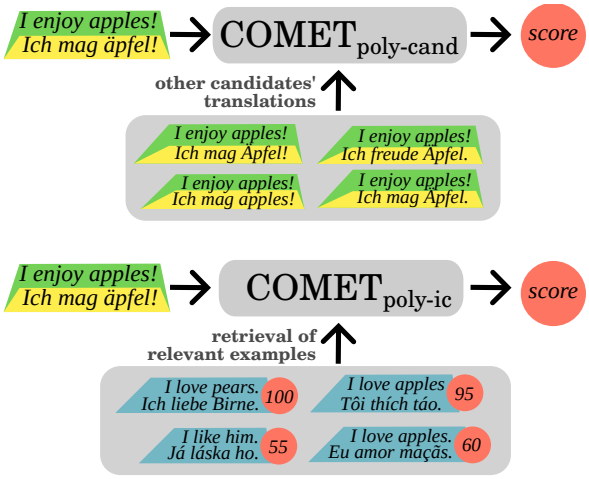


Figure 1: The $\text{COMET}_{\text{poly-ic}}$ model consults a knowledge base of previously human-scored translations before assigning the quality estimation score to the candidate translation. The $\text{COMET}_{\text{poly-cand}}$ considers other possible translations apart from the candidate one. Both metrics work better than just providing the source and the translation.

1 Introduction

There is a gap between how humans and automated metrics score translations. Automated metrics receive the source segment, usually a sentence or a paragraph, a single translation, and optionally a reference translation. They are then tasked with assessing the quality of the translation. In contrast, human evaluation is less episodic. Human raters often assess multiple translations in sequence (Graham et al., 2013; Freitag et al., 2021; Kocmi et al., 2024b), considering them side-by-side. Even though annotations are made for each translation individually, annotators become calibrated (known as sequence effect, Mathur et al., 2017), to common error patterns and their own evaluation criteria as they review multiple translations. As a result, they effectively score each translation in the context of others. Moreover, unlike human annotators, who have a deep understanding of the languages involved and can assess a wide range of translation qualities, automated metrics are limited by the data they were trained on. As a result, their performance tends to degrade when evaluating translations that deviate from their training distributions, such as out-of-domain content (Zouhar et al., 2024).

We present two conceptual approaches to address these two challenges by incorporating additional context into standard automated metrics, such as COMET (Rei et al., 2020). Our main motivation is to narrow the gap between human evaluation and automated metrics, enabling automated metrics to score translations in the context of other

★Equal contribution, sorted anti-alphabetically.

[1]We release the paper code and pre-trained quality estimation models $\text{COMET}_{\text{poly-ic}}$ and $\text{COMET}_{\text{poly-cand}}$.

Proceedings of the Tenth Conference on Machine Translation, pages 887–904
November 8-9, 2025

translations and making them more robust to out-of-domain data. Specifically, we introduce two models trained within this framework, $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$:

- In $\text{COMET}_{\text{poly-cand}}$, different translations of the same source sentence are provided to the model as additional context (Figure 1 top). This is suitable for scenarios such as (1) benchmarking, where we evaluate translations of multiple systems on the same source sentence, or (2) reranking, where we need to select the best translation from a pool of candidate translations.
- In $\text{COMET}_{\text{poly-ic}}$, which is inspired by retrieval-based in-context learning, tuples of (*source*, *translation*, *human quality score*) are provided to the model as additional context. The tuples are retrieved based on the source sentence similarity to the evaluation example at hand (Figure 1 bottom). In practice, in-context examples can be obtained from existing, previously scored translations—such as those found in prior WMT annotation datasets (Freitag et al., 2024; Kocmi et al., 2024a).

This paper is structured as follows. In Section 2, we first describe the task of machine translation quality estimation (QE) and COMET (Rei et al., 2020), a popular QE metric. Then, we describe our two proposed model variants, $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$. We also apply the same approach to two additional QE systems with contrasting characteristics: a non-parametric k-NN baseline and GEMBA, a large parametric LLM-based evaluator. In Sections 3 and 4, we show that our methods not only improve COMET's segment-level performance but also outperform both the much larger GEMBA model and the k-NN baseline, despite their simplicity. These approaches also show promise for instant on-the-fly domain adaptation. We place our contributions in context with related work in Section 5. Finally, in Section 6, we provide some practical guidance on using these metrics along with potential caveats.

We publicly release our models under open license, and submit our models to the WMT 2025 Metrics Shared Task.

2 Methods

In this section, we introduce the translation quality estimation task, review COMET, and present two extensions for improved quality estimation and domain adaptation.

2.1 Background

Quality estimation (QE). Given a source text s and a model-produced translation (MT) t, which is assessed by a human annotator on a scale from 0% to 100%, the goal of quality estimation (QE) is to develop a metric to predict this score.

Baseline COMET. Traditionally, quality estimation relied on static, rule-based metrics, but the field has shifted toward learned, data-driven metrics that can better approximate human judgments (Freitag et al., 2022). Learned automated metrics can be thought of as a function f, taking a source sentence s and a translation t as input and producing a continuous score $f(s,t) \in [0,1]$. f is usually trained in a supervised manner to approximate human judgment $y_{s,t} = \text{human}(s,t)$:

$$f(s,t) \xrightarrow{\text{train}} y_{s,t}$$

A popular recent choice for f is COMET (Rei et al., 2020), which is a combination of a trainable encoder model e_{θ_1} and a multi-layer perception head MLP_{θ_2}. COMET first embeds the source and translation texts, obtaining $s^e = e_{\theta_1}(s)$ and $t^e = e_{\theta_1}(t)$, and then transforms the embeddings into a score prediction using MLP_{θ_2}. We denote the set of trainable weights as $\theta = (\theta_1, \theta_2)$. Specifically, COMET is formulated as:

$$\text{COMET}_\theta(s,t) = \text{MLP}_{\theta_2}\big(g_{\theta_1}(s,t)\big),$$
$$\text{with } g_{\theta_1}(s,t) = \langle s^e, t^e, |s^e - t^e|, s^e * t^e \rangle.$$

Here, g_{θ_1} constructs a feature vector for the pair (s,t) by concatenating their embeddings s^e and t^e with additional element-wise transformations: the absolute difference $|s^e - t^e|$ and the element-wise product $s^e * t^e$. The trainable weights θ are optimized by minimizing the mean squared error between the COMET score and human labels using a variation of the stochastic gradient descent algorithm.

While the baseline COMET framework is effective, it does not support incorporating additional information. Just like for human evaluators, having more information such as (1) multiple candidates' translations for the same source, or (2) ground-truth example quality scores of translations, could improve the performance of COMET further. Thus, we introduce two extensions for COMET, which we train from scratch..

2.2 Multiple Candidates: $\text{COMET}_{\text{poly-cand}}$

Our first variant targets scenarios like benchmarking or reranking MT models, where multiple translations of the same source segment are available. In these cases, we extend the model's context by including additional translations $\{t_i\}_{i=2}^{n}$ of the same source sentence s, allowing the model to leverage multiple candidate translations simultaneously. Figure 1 (top) shows an illustration of this model architecture.

Specifically, we include the embeddings of these additional translations as part of the input to the multi-layer perceptron. Formally, for all $i \in \{2, \cdots, n\}$, we define

$$g_{\theta_1}(t, t_i) = \langle t_i^e, |t_i^e - t^e|, t_i^e * t^e \rangle.$$

We then concatenate $\langle g_{\theta_1}(s,t), g_{\theta_1}(t,t_1), ..., g_{\theta_1}(t,t_n) \rangle$ and pass it to the MLP. During training, we ensure that the additional translations $\{t_i\}_{i=2}^{n}$ differ from the main translation t, and keep n fixed across all training examples.

Joint predictions. To reduce computation time, $\text{COMET}_{\text{poly-cand}}$ can be trained to jointly predict the quality scores of the original translation along with those of the additional translations. The training objective then becomes:

$$f(s, t, t_2, ..., t_n) \xrightarrow{\text{train}} y_{s,t}, y_{s,t_2}, ..., y_{s,t_n}$$

Using scores of other translations. When the human assessment scores for additional translations, $\{y_{s,t_i}\}_{i=2}^{n}$, are available, we can further augment the feature vector using these scores. The input to the MLP would become:

$$\langle g_{\theta_1}(s,t), g_{\theta_1}(t,t_1), y_{s,t_1}, ..., g_{\theta_1}(t,t_n), y_{s,t_n} \rangle$$

This is particularly useful when we wish to evaluate a new system on a pre-existing benchmark with other candidate translations whose qualities are already annotated by humans.

2.3 In-context Learning: $\text{COMET}_{\text{poly-ic}}$

In the previous approach, we used additional translations of the same source sentence, a setup that might be unrealistic outside controlled scenarios such as benchmarking or reranking. An alternative, inspired by the success of *in-context* learning in other domains (Brown et al., 2020), is to provide the model with other, similar examples: by conditioning on human-scored translations, it can learn the mapping between translation patterns and quality judgments on the fly.

$\text{COMET}_{\text{poly-ic}}$ implements this by retrieving source–translation–score triplets from a knowledge base, in our case, prior WMT annotation datasets (Freitag et al., 2024; Kocmi et al., 2024a), and using them as context, enabling the model to adapt its evaluation to different domains. An illustration is shown in Figure 1 (bottom).

Specifically, for each input example (source s, translation t), we retrieve the examples $\{(s_i, t_i, y_{s_i,t_i})\}_{i=2}^{n_{\text{ICL}}+1}$ from a knowledge base $\mathcal{D}$. The new examples are added to the representation vector similar to $\text{COMET}_{\text{poly-cand}}$, considering both embeddings and labels, by appending

$$\langle t_i^e, |t_i^e - t^e|, t_i^e * t^e, s_i^e, |s_i^e - s^e|, |s_i^e * s^e|, y_{s_i,t_i} \rangle$$

to $g_{\theta_1}(s,t)$ for all $i \in \{2, \cdots, n_{\text{ICL}} + 1\}$.

The ICL examples are retrieved using normalized embedding (cosine) similarity computed from either the source s^e (default), the translation t_i^e, their arithmetic combination $s^e + t_i^e$, or their concatenation $\langle s^e, t_i^e \rangle$. We retrieve up to five most similar examples, discarding exact matches during training. We present detailed ablations of different filtering and retrieval setups in Section 4.

2.4 Including Reference Translations

Optionally, COMET can also make use of a reference translation (Rei et al., 2020), though this is no longer part of the standard QE setup. We also report results for $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$ in the reference-based setting, by incorporating the reference r in their inputs, i.e., $f(s, t, r)$. However, our primary focus remains on QE, as references are often unavailable in practical scenarios.

2.5 Models Beyond COMET

Since our method is not specific to COMET, we include two models that, like our extensions, can take multiple candidate translations into account.

***k*-nearest neighbors.** As our first baseline method, we propose using a k-nearest-neighbours (k-NN) approach, mirroring methods used in similar contexts (Dinh et al., 2024). k-NN naturally leverages existing high-quality examples by retrieving similar instances, providing a strong non-parametric baseline that complements our model-based approaches. The k-NN baseline is implemented for our two different setups: $k\text{-NN}_{\text{poly-cand}}$ and $k\text{-NN}_{\text{poly-ic}}$.

For k-NN$_{\text{poly-cand}}$ and for a pair (s, t), we retrieve k additional translations for s. These are selected based on the cosine similarity of the target translation t and candidate translations, where embeddings are obtained using the all-MiniLM-L12-v2 (Reimers and Gurevych, 2020), yielding the set $\{t_i\}_{i=2}^{k+1}$. We then rate each candidate using Baseline COMET, obtaining $\{\text{COMET}(s, t_i)\}_{i=2}^{k+1}$. Finally, we use their average as the final prediction, i.e.,

$$\hat{y}_{s,t}^{k\text{-NN}_{\text{poly-cand}}} = \frac{1}{k}\sum_{i=2}^{k+1}\text{COMET}(s, t_i).$$

For k-NN$_{\text{poly-ic}}$, we retrieve $k = n_{\text{ICL}}$ examples, $\{(s_i, t_i, y_{s_i,t_i})\}_{i=2}^{k+1}$, following the retrieval strategies described in Section 2.3, and then average their human scores to obtain the prediction:

$$\hat{y}_{s,t}^{k\text{-NN}_{\text{poly-ic}}} = \frac{1}{k}\sum_{i=2}^{k+1} y_{s_i,t_i}.$$

We further extend the k-NN approaches using weighted averages in Appendix D.

Using LLMs as evaluators. As a second baseline, we use large language models (LLMs) for MT evaluation, leveraging their effectiveness in this task (Kocmi and Federmann, 2023). Specifically, we apply in-context learning (Brown et al., 2020), a standard method for injecting new knowledge into LLMs at inference time. Similar to COMET$_{\text{poly}}$, we provide LLMs with additional contextual information when scoring translations. However, unlike COMET$_{\text{poly}}$ variants, which update model parameters during training, LLMs receive this information only through their prompts at inference time, without any parameter modification.

For prompt creation, we build on top of GEMBA (Kocmi and Federmann, 2023), a framework designed to prompt LLMs to score the quality of translations. Leveraging GEMBA's pre-defined prompts, we extend them to two settings: (1) GEMBA$_{\text{poly-cand}}$, where additional translations of the same source sentence are provided, and (2) GEMBA$_{\text{poly-ic}}$, where full examples (including source, translation, and human quality score) are included. Prompt details are provided in Appendix A.2.

3 Experimental Setup

This section outlines the training and evaluation procedures, as well as the experimental setup.

Data. We use the direct assessment scorings of WMT up to 2023 (inclusive) for training (600k segments). For testing and evaluation, we use WMT 2024 (105 segments), which has been evaluated with the ESA protocol (Kocmi et al., 2024b). This dataset covers eleven language pairs: English to Czech, German, Spanish, Hindi, Icelandic, Japanese, Russian, Chinese, Czech to Ukrainian, and Japanese to Chinese. From ESA, we use the final scores (as opposed to error spans), which have the same scale as direct assessment. For MQM, we convert the error span annotations on a translation to the final score by taking $1 - (5 \cdot major + 1 \cdot minor)/100$, where $major$ is the number of annotated major errors, and $minor$ is the number of minor errors annotated in the translation. In this way, the scores are aligned roughly on the same scale compared to DA scores.

Training. We train the Baseline COMET model, COMET$_{\text{poly-cand}}$ and COMET$_{\text{poly-ic}}$ based on pretrained RoBERTa (Liu et al., 2019) on WMT human judgment data for five epochs. For COMET$_{\text{poly-cand}}$, we retrieve up to five candidate translations, either randomly or based on embedding similarity. For COMET$_{\text{poly-ic}}$, we retrieve up to five in-context examples from the training data based on embedding similarity. The metrics are trained in a maximally comparable model setup, which is detailed in Appendix A.

Evaluation. We evaluate the metrics on the segment level in three ways: Pearson correlation, Kendall's tau-b, and Mean Absolute Error (MAE). In contrast to Freitag et al. (2024) we do not do perform any group-by-item nor group-by-item. Results are macro-averaged across eleven languages.

Pearson correlation measures the linear relationship between metric scores and human ratings: higher values indicate better alignment, though not necessarily on the same scale. Mean Absolute Error (MAE), in contrast, captures the average absolute difference between metric and human scores, with lower values indicating closer agreement in both value and scale. Kendall's tau-b focuses on rank correlation, reflecting how well the metric preserves the relative ordering of translations. While Pearson and Kendall's tau-b range from -1 to 1, MAE is unbounded and depends on the scoring scale.

Experiments. To ensure a controlled evaluation setting, we first train a standard COMET

Model	Reference-less $\rho\uparrow$	Reference-less $\tau_b\uparrow$	Reference-less MAE $\downarrow$	Reference-based $\rho\uparrow$	Reference-based $\tau_b\uparrow$	Reference-based MAE $\downarrow$
standard COMET model $f(s,t) \rightarrow \hat{y}_t$	0.105	0.079	30.2	0.245	0.166	26.6
COMET$_{\text{poly-cand}}$						
additional candidate $f(s,t,t_2^*) \rightarrow \hat{y}_t$	0.160	0.127	28.5	0.281	0.180	26.3
additional candidate, output joint predictions $f(s,t,t_2^*) \rightarrow \hat{y}_t, \hat{y_{t_2^*}}$	0.167	0.113	28.8	0.275	0.172	25.6
additional candidate and its score $f(s,t,t_2^*,y_{t^*2}) \rightarrow \hat{y}_t$	0.267	0.207	21.9	0.374	0.243	20.6
COMET$_{\text{poly-ic}}$						
additional candidate and its score $f(s,t,t_2^*,y_{t^*2}) \rightarrow \hat{y}_t$	0.141	0.116	27.3	0.352	0.247	15.3

Table 1: Results for COMET$_{\text{poly-cand}}$ and COMET$_{\text{poly-ic}}$. The first row shows the standard COMET. The middle and bottom parts show that adding additional translation candidates and in-context examples boosts performance.

Model (Reference-less)	$\rho\uparrow$					$\tau_b\uparrow$					MAE $\downarrow$				
(+additional)	+1	+2	+3	+4	+5	+1	+2	+3	+4	+5	+1	+2	+3	+4	+5
$f(s,t) \rightarrow \hat{y}_t$	0.105	0.105	0.105	0.105	0.105	0.079	0.079	0.079	0.079	0.079	30.2	30.2	30.2	30.2	30.2
COMET$_{\text{poly-cand}}$															
$f(s,t,t_{...}) \rightarrow \hat{y}_t$	0.160	0.251	0.224	0.202	0.190	0.127	0.145	0.127	0.130	0.120	28.5	27.6	26.8	28.4	27.6
$f(s,t,t_{...},y_{t_{...}}) \rightarrow \hat{y}_t$	0.267	0.321	0.328	0.327	0.321	0.207	0.229	0.230	0.235	0.233	21.9	17.3	16.0	14.0	13.7
COMET$_{\text{poly-ic}}$															
$f(s,t,t_{...},y_{t_{...}}) \rightarrow \hat{y}_t$	0.141	0.134	0.148	0.128	0.068	0.116	0.108	0.114	0.105	0.075	27.3	27.2	24.7	27.6	27.4

Table 2: Results for COMET$_{\text{poly-cand}}$ and COMET$_{\text{poly-ic}}$ using different numbers of additional translation candidates. The +1 is equal to the results in Table 1. The +x uses x additional translation candidates, which improves performance especially for COMET$_{\text{poly-cand}}$.

model on the data described before and use it as a baseline. We then investigate COMET$_{\text{poly-cand}}$ by incorporating additional translations into the base model and analysing the impact of different selection strategies. Similarly, we explore COMET$_{\text{poly-ic}}$, experimenting with various retrieval methods and assessing its potential for domain adaptation. We complement our experiments with k-NN$_{\text{poly-cand}}$ and k-NN$_{\text{poly-ic}}$ as non-parametric baselines, and GEMBA$_{\text{poly-cand}}$ and GEMBA$_{\text{poly-ic}}$ as large-parameter LLM baselines.

4 Results and Analysis

In the following, we discuss and analyse the results of COMET$_{\text{poly-cand}}$ and COMET$_{\text{poly-ic}}$, compare them to the non-parametric k-NN and the large parametric GEMBA model, and discuss the runtime impact of our method.

4.1 Results for COMET$_{\text{poly-cand}}$

Additional candidate helps. We begin by evaluating COMET$_{\text{poly-cand}}$ in its simplest setting: adding a single additional translation from the same source as the candidate being scored. We choose the closest additional translation t_2^* as, intuitively, the closer it is to the candidate t, the more relevant it is for assessing its quality. We select t_2^* based on the embedding distance computed between candidate translations (see Appendix A for details on embeddings and distance metrics). The corresponding results are shown in the middle part of Table 1.

Across all evaluation metrics, including an additional translation $f(s,t,t_2)$, considerably improves performance compared to the standard COMET baseline $f(s,t)$. Specifically, Pearson correlation improved by over 50%. The joint translation prediction objective, which scores both the original translation and the additional translation, also yields gains over the baseline, though it performs slightly worse than the single-prediction setup. This suggests that, in scenarios where faster inference is needed, the joint-prediction setup offers a practical trade-off, delivering improved performance with smaller additional cost. Finally, including the gold score y_{t_2} of the additional translation in the input vastly improves the metric performance. However, note that this is an ideal scenario where the gold score y_{t_2} is available, which is not always realistic.

Note that it is not always possible to find additional translations that are similar to the translation

at hand. Therefore, we experiment with using a randomly selected additional candidate to test the robustness of $\text{COMET}_{\text{poly-cand}}$. This still results in notable gains, albeit smaller than with similar candidates. We report these results in Appendix B.

More than one candidate helps. We extend $\text{COMET}_{\text{poly-cand}}$ by increasing the number of additional candidates. The results are shown in Table 2. Having more than one additional candidate further improves the performance of $\text{COMET}_{\text{poly-cand}}$, as we are providing a more global view of possible translations to the model. However, this effect starts to diminish beyond two additional candidates. For comparison, results using random additional candidates are provided in Appendix B.

Additional translation complement reference. Previous experiments focused on reference-free evaluation. To complete the picture, we now explore how $\text{COMET}_{\text{poly-cand}}$ performs when reference translations are available.

The right half of Table 1 shows that using $\text{COMET}_{\text{poly-cand}}$ with reference yields better performance than $\text{COMET}_{\text{poly-cand}}$ in QE mode, though the gain is smaller than for standard COMET. This indicates that additional translations help narrow the gap but cannot fully replace references. Rather, additional translations complement references by providing further improvements on top of them.

4.2 Results for $\text{COMET}_{\text{poly-ic}}$

Building on this idea of leveraging additional context, we next evaluate $\text{COMET}_{\text{poly-ic}}$, which incorporates in-context examples to further enhance evaluation quality.

In-context examples help. We retrieve an in-context example using the source text s^e, embedded via an external embedding model (details in Appendix C). Results in the bottom row of Table 1 show that COMET benefits significantly from these examples, outperforming the baseline without in-context examples. This improvement also holds for $\text{COMET}_{\text{poly-ic}}$ with references. However, compared to $\text{COMET}_{\text{poly-cand}}$, in-context examples appear less informative than additional candidates with the same source, resulting in slightly reduced performance. We also test other embedding types (including COMET's own) and variations using the target or both source and target for retrieval. However, none of these alternatives yields further improvements. Full ablations are presented in Appendix C.

(a) k-$\text{NN}_{\text{poly-cand}}$

k	$\rho\uparrow$	$\tau_b\uparrow$	MAE$\downarrow$
1	0.083	0.064	30.4
2	0.087	0.064	30.3
3	0.086	0.062	30.4
4	0.085	0.059	30.4
5	0.085	0.057	30.4

(b) k-$\text{NN}_{\text{poly-ic}}$

k	$\rho\uparrow$	$\tau_b\uparrow$	MAE$\downarrow$
1	0.029	0.014	31.1
2	0.031	0.017	29.4
3	0.034	0.017	28.7
4	0.036	0.019	28.2
5	0.037	0.020	27.9

Table 3: Results for the k-nearest neighbors baseline using embeddings $\langle s^e, t_i^e \rangle$ in both k-$\text{NN}_{\text{poly-cand}}$ and k-$\text{NN}_{\text{poly-ic}}$ setup. k-NN consistently underperforms $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$, showing notably lower correlations despite comparable MAE.

More in-context examples improve performance. While a single in-context example already boosts performance, adding up to three examples leads to further improvements. As shown in the bottom half of Table 2, performance increases with the number of retrieved examples using the external embedding model and s^e for retrieval, but declines beyond three examples, likely because additional examples become less similar and less relevant.

We also provide preliminary experiments in Appendix C.3 on how $\text{COMET}_{\text{poly-ic}}$ can leverage in-context examples to adapt its quality estimation to a new domain, and find a slight improvement compared to the base model.

4.3 Adding Candidates to Models Beyond COMET

In order to see whether having additional candidates or examples also helps with other QE methods other than COMET, we look into the performance of two baselines: the non-parametric ***k***-nearest neighbors and large parametric LLM evaluator with GEMBA.

We use ***k***-nearest neighbors in the retrieval setting for both k-$\text{NN}_{\text{poly-cand}}$ and k-$\text{NN}_{\text{poly-ic}}$, i.e., retrieving similar examples along with their gold quality scores, since the gold scores are required for ***k***-nearest neighbors. For GEMBA, we experiment with all $\text{GEMBA}_{\text{poly-cand}}$ variances (random/similar candidate, with/without gold scores) and $\text{GEMBA}_{\text{poly-ic}}$, similar to $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$.

***k*-nearest neighbors underperforms COMET.** We present results for the k-nearest neighbors (k-NN) baseline in Table 3, varying k from 1 to 5,

	Input → Output	Reference-less			Reference-based		
		$\rho \uparrow$	$\tau_b \uparrow$	MAE $\downarrow$	$\rho \uparrow$	$\tau_b \uparrow$	MAE $\downarrow$
standard GEMBA	$f(s,t) \rightarrow \hat{y}_t$	0.266	0.199	27.6	0.311	0.200	27.3
GEMBA$_{\text{poly-cand}}$, closest t_2^*							
additional candidate	$f(s,t,t_2^*) \rightarrow \hat{y}_t$	0.245	0.185	28.2	0.277	0.187	27.5
additional candidate, joint predictions	$f(s,t,t_2^*) \rightarrow \hat{y}_t, \hat{y_{t_2^*}}$	0.235	0.149	28.6	0.296	0.181	27.9
additional candidate and its score	$f(s,t,t_2^*,y_{t^*2}) \rightarrow \hat{y}_t$	0.276	0.187	27.4	0.337	0.217	26.8
GEMBA$_{\text{poly-ic}}$							
additional candidate and its score	$f(s,t,s_2,t_2,y_{t_2}) \rightarrow \hat{y}_t$	0.195	0.099	28.3	0.291	0.168	27.4

Table 4: Results for GEMBA$_{\text{poly-cand}}$ and GEMBA$_{\text{poly-ic}}$. The first row shows the standard GEMBA model. In contrast to the COMET models, adding additional translation candidates and in-context examples does not significantly boost performance.

along with the simple average approach. For k-NN$_{\text{poly-ic}}$, neighbors are retrieved using the embedding $\langle s^e, t_i^e \rangle$. k-NN$_{\text{poly-ic}}$ performs markedly worse than our COMET variants (COMET$_{\text{poly-cand}}$ and COMET$_{\text{poly-ic}}$), particularly on correlation metrics, though MAE differences remain small. This is expected, as ***k***-nearest neighbors naively aggregate the scores of the closest datapoints, without actually modeling the underlying relationships between the source and translation to output the quality score. In the cases where the neighbors are not close enough, the output from ***k***-nearest neighbors would be suboptimal. In the poly-cand scenario, k-NN$_{\text{poly-cand}}$ achieves results similar to the naive COMET approach, unsurprising given that k-NN in this case effectively averages COMET scores for similar translations.

A more comprehensive set of results is provided in Appendix D, including a weighted variant of the ***k***-nearest neighbors baseline. The appendix also compares different retrieval strategies for k-NN$_{\text{poly-ic}}$. Among them, retrieval using $\langle s^e, t_i^e \rangle$ performs best; this contrasts with COMET$_{\text{poly-ic}}$, where retrieving based solely on the source yields better results. This difference arises because retrieval based only on source can hurt k-NN$_{\text{poly-ic}}$ by averaging scores from translations that may not align well with the target one.

COMET$_{\text{poly-cand}}$ outperforms GEMBA. We now move on to the parameter-heavy LLM baseline GEMBA. The main results for GEMBA$_{\text{poly-cand}}$ and GEMBA$_{\text{poly-ic}}$ are shown in Table 4.

Due to the large size and large amount of pretraining data of LLMs, the baseline GEMBA model has notably better performance than the baseline COMET (0.266 Pearson versus 0.105 Pearson). However, GEMBA does not benefit from our poly-cand and poly-ic setup. In most configurations, neither method improves over the baseline. Consequently, by better making use of additional examples, the COMET$_{\text{poly-cand}}$ variance outperforms all GEMBA variances. The exception is GEMBA$_{\text{poly-cand}}$ with the closest additional translation and its gold quality score, which yields better performance than baseline GEMBA. This is unsurprising, as the target translation's quality is likely similar to that of its closest neighbor, whose score is provided to the model. We also test adding random or multiple examples; random candidates perform comparably to similar ones, while multiple examples do not consistently yield further gains. Detailed results can be found in Appendix E.

4.4 Comparing Efficiency of COMET-poly Models

While the previous section shows that COMET$_{\text{poly-cand}}$ outperforms GEMBA in certain evaluation settings, this advantage is even more significant in practice due to efficiency. Table 5 shows that overall, running GEMBA is considerably slower and requires more computational resources than COMET. This highlights the benefits of training a small, specialized model (COMET$_{\text{poly-cand}}$) to match the performance of large, general-purpose models (GEMBA), while substantially reducing inference-time computational costs.

On the other hand, compared to k-NN, COMET$_{\text{poly}}$ is less efficient. k-NN is non-parametric, thus its computation time is almost instantaneous when excluding retrieval cost. However, as we have seen in the previous section, k-NN has notably worse performance compared to COMET$_{\text{poly}}$.

We next examine the general runtime behavior

	COMET	GEMBA
standard model		
$f(s,t) \rightarrow \hat{y_t}$	4.4s/1k	196.1s/1k
poly-cand		
$f(s,t,t_2) \rightarrow \hat{y_t}$	6.9s/1k	254.0s/1k
$f(s,t,t_2) \rightarrow \hat{y_t}, \hat{y_{t_2}}$	3.5s/1k	146.3s/1k
$f(s,t,t_2,y_{t_2}) \rightarrow \hat{y_t}$	6.9s/1k	256.0s/1k
poly-ic		
$f(s,t,s_2,t_2,y_{t_2}) \rightarrow \hat{y_t}$	7.2s/1k	233.0s/1k

Table 5: Inference time of GEMBA models compared to COMET models on the WMT 2024 test set (time per 1000 scores output on a single NVIDIA H100). COMET has ∼0.5B params and GEMBA 70B. GEMBA is run with 4-bit quantization. $\text{COMET}_{\text{poly-ic}}$ introduces an additional cost of retrieving from a vector knowledge base which we exclude for both $\text{COMET}_{\text{poly-ic}}$ and $\text{GEMBA}_{\text{poly-ic}}$.

of our methods across multiple settings. Looking at Table 5, unsurprisingly, integrating additional candidates $f(s,t,t_2)$ is more expensive in comparison to the baseline model with only one translation $f(s,t)$. However, most of the computation is spent on encoding the text sequences, which can be efficiently cached during inference (Rei et al., 2022), making all of the metric variations comparable. Moreover, if both t and t_2 need to be scored, then using a model that predicts both of their scores $\hat{y_t}, \hat{y_{t_2}}$ is faster than computing $f(s,t)$ and $f(s,t_2)$ together.

4.5 Analysis

To better understand the impact of our method, we investigate how additional translations or samples influence COMET's quality predictions.

$\text{COMET}_{\text{poly-cand}}$. We first perform a systematic analysis by categorizing test cases according to the gold quality scores of both the translation under evaluation and its additional translation. Specifically, we consider four combinations: (i) both high-quality, (ii) sample high / additional low, (iii) sample low / additional high, and (iv) both low-quality.

Results show that additional translations are most beneficial when the evaluated output is of lower quality. Interestingly, the quality of the additional translation itself has little impact on QE performance. This suggests that even low-quality additions can aid COMET by introducing complementary error patterns that highlight discrepancies. Detailed results can be found in Appendix F.

We then focus on individual cases where the additional translation yields the largest improvements. To do so, we sort the test samples in descending order by the difference between COMET's absolute error and that of $\text{COMET}_{\text{poly-cand}}$, thereby identifying the samples where $\text{COMET}_{\text{poly-cand}}$ yields the greatest improvement. We then conduct a manual inspection of the top cases, revealing that additional translations help COMET better detect specific failure modes: undertranslation, where the translation is merely a copy of the source; numerical errors, where numeric values in the translation differ from the source; explanations, where unnecessary explanatory text is added; and refusals, where the translation includes statements declining to translate the input. In these cases, the additional translations do not exhibit the same errors as the translation under evaluation. We therefore hypothesize that the additional translations effectively serve as references in such scenarios. We provide specific examples in Appendix B.2 in Appendix F.

$\text{COMET}_{\text{poly-ic}}$. We perform a similar systematic analysis for $\text{COMET}_{\text{poly-ic}}$ to study how in-context examples influence the scoring of high- and low-quality translations. Consistent with $\text{COMET}_{\text{poly-cand}}$, $\text{COMET}_{\text{poly-ic}}$ shows greater benefits when evaluating lower-quality outputs (see Appendix F for details).

In addition, we also investigate the choice of in-context examples, which is critical for $\text{COMET}_{\text{poly-ic}}$'s performance. During training, retrieved examples are drawn from the training set and thus come from the same distribution and have been seen by the model. In contrast, at test time, the examples are unseen and often less similar. We investigate whether the train-test mismatch affects $\text{COMET}_{\text{poly-ic}}$ by training models with different similarity thresholds. However, we find that the train-test mismatch does not significantly impact performance. Details can be found in Appendix C.4.

5 Related Work

This section reviews the broader context of automated metrics and human evaluations that use multiple inputs: either multiple translations or, more commonly, multiple references.

Automated metrics. Early metrics like BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) operate at segment or corpus level and support multiple references but not multiple hypotheses simultaneously. COMET (Rei et al., 2020) trains an

encoder for human-like quality assessment and supports a single reference. Adding more references shows limited gains (Zouhar and Bojar, 2024).

Closest to our work, Dinh et al. (2024) propose a k-NN quality estimator similar to $\text{COMET}_{\text{poly-ic}}$, but aggregate train-test similarity of MT models as a quality indicator rather than having a separate QE model that assesses translations based on similarity and contextual relevance. Moosa et al. (2024) introduce MT-Ranker, which compares translation pairs and outputs a binary preference.

With the rise of Large Language Models (LLMs), an up-to-date approach for Quality Estimation is to use LLM-as-a-Judge. Simply prompting LLMs to output the quality score of a translation has become the state-of-the-art approach, with the most prominent example of GEMBA (Kocmi and Federmann, 2023). This approach has the potential to improve even further, by applying different strategies such as including in-context examples (few-shot judge), chain-of-thought prompting, pairwise comparison, as recommended by Zheng et al. (2023).

Human evaluation. Human evaluation of machine translation takes many forms. For benchmarking, WMT initially used RankME (Novikova et al., 2018), where annotators rank multiple hypotheses simultaneously.

Due to biases and high cognitive load, this shifted to single-hypothesis assessments such as Direct Assessment and its variants (Graham et al., 2013; Kocmi et al., 2022), Multidimensional Quality Metrics (Freitag et al., 2021), and Error Span Annotation and its variants (Kocmi et al., 2024b; Zouhar et al., 2025). Despite judging one hypothesis at a time, annotators gradually see other translations during evaluation, implicitly calibrating their quality judgments. Automated metrics, however, lack this contextual grounding and evaluate translations independently.

6 Discussion and Conclusion

Recommendation. $\text{COMET}_{\text{poly-cand}}$ can be applied in scenarios where multiple translations exist for the same source sentence, such as: (1) enchmarking various competing systems on the same test set (e.g., WMT General shared tasks), (2) comparing outputs from different checkpoints or models during MT development, or (3) cselecting the best translation from a pool of hypotheses during reranking for final output selection.

The intended use of $\text{COMET}_{\text{poly-ic}}$ is for quick domain adaptation without retraining the metric (Appendix C.3). While different retrieval methods can cause slight variations in performance (see Appendix C), it is crucial that the retrieval mechanism is deterministic to ensure reproducible scores. Additionally, changing the retrieval mechanism or the set of previously annotated translations that are being retrieved instantiates a new metric with non-comparable scores to the previous evaluations. Therefore, when using $\text{COMET}_{\text{poly-ic}}$, always disclose the retrieval set and retrieval method.

Training a smaller, specialized module with some tweaks ($\text{COMET}_{\text{poly-cand}}$) can be beneficial compared to directly using large, general-purpose language models (GEMBA). We have shown that $\text{COMET}_{\text{poly-cand}}$ can reach the performance of GEMBA, while being much more efficient in terms of inference time.

Submitted models. We submit the following models to the WMT Metrics Shared Task 2025 and make them publicly available under open license (Apache License 2.0) on Hugging Face. The models are trained on WMT data up to 2024 (inclusive).

- COMET-poly-base-wmt25: baseline
- COMET-poly-cand1-wmt25: one additional translation
- COMET-poly-cand2-wmt25: two additional translations
- COMET-poly-ic1-wmt25: one in-context example
- COMET-poly-ic3-wmt25: three in-context examples
- knn-poly-cand3: three additional translations, scored with COMET-poly-base-wmt25
- knn-poly-ic3: three in-context examples

Conclusion. In this work, we introduced two new paradigms for machine translation quality estimation: (1) evaluating a translation with the context of other translations of the same source, and (2) quality estimation with retrieval for in-context examples. We showed that these approaches show potential in being more adaptable and outperforming the baseline COMET, while also offering practical advantages in efficiency by matching the performance of larger models at lower computational cost.

Limitations

$\text{COMET}_{\text{poly-cand}}$ is entirely constrained to setups where we are scoring multiple translations at the same time. This is by design and thus mostly suited for WMT-style benchmarking competitions or model development where we wish to find which translation model is the best one. It is not useful for scenarios where a single model is being evaluated without the context of other existing translations.

Both $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$ are not exempt on the reliance on the quality of previously human-annotated translations. In some cases, the quality of the collected data might be subpar (Kocmi et al., 2024a), which is then then further exemplifies its bias in $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$.

Our investigation in this paper omits various tricks used to further boost COMET's performance for the purpose of clarity of the core methodological contributions of $\text{COMET}_{\text{poly-cand}}$ and $\text{COMET}_{\text{poly-ic}}$.

Ethics Statement

Vilém Zouhar declares a potential conflict of interest as an organizer of the WMT 2025 Metrics Shared Task. No privileged information has been used in this work.

Acknowledgements

This research has been funded in part by a Swiss National Science Foundation award (project 201009) and a Responsible AI grant by the Haslerstiftung. Part of this work received support from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People). This work was also supported by the Helmholtz Programme-oriented Funding, with project number 46.24.01, project name AI for Language Technologies. We acknowledge the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Wurttemberg and by the Federal Ministry of Education and Research.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Tu Anh Dinh, Tobias Palzer, and Jan Niehues. 2024. Quality estimation with k-nearest neighbors and automatic evaluation for model-specific quality estimation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, 133–146, Sheffield, UK. European Association for Machine Translation (EAMT).

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, 47–81. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 46–68. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 33–41. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet.

In *Proceedings of the Ninth Conference on Machine Translation*, 1–46. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 1–45. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, 193–203. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, 1440–1453. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2017. Sequence effects in crowdsourced annotations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2860–2865. Association for Computational Linguistics.

Ibraheem Muhammad Moosa, Rui Zhang, and Wenpeng Yin. 2024. MT-ranker: Reference-free machine translation evaluation by inter-system ranking. In *The Twelfth International Conference on Learning Representations*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 72–78. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 61–70. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2685–2702. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4512–4525. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Vilém Zouhar and Ondřej Bojar. 2024. Quality and quantity of machine translation references for automatic metrics. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, 1–11. ELRA and ICCL.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 488–500. Association for Computational Linguistics.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Appendix Overview

The appendix includes the following information:

- Implementation Details (§A)
- $\text{COMET}_{\text{poly-cand}}$ Ablations and Analysis (§B)
- $\text{COMET}_{\text{poly-ic}}$ Ablations and Analysis (§C)
- k-NN Ablations and Analysis (§D)
- GEMBA Ablations and Analysis (§E)
- Analysis of Impact of Additional Translations and In-Context Examples (§F)

A Implementation details

A.1 COMET

The model details are shown in Table 6. For computing embeddings to retrieve similar examples, by default we use the cosine distance from all-MiniLM-L12-v2 (Reimers and Gurevych, 2020). However, we also experiment in ablations with using the xlm-roberta-large (Conneau et al., 2020) embeddings and embeddings from a trained baseline COMET.

Encoder	xlm-roberta-large (24 layers)
Embeddings	Layerwise attention & CLS
Encoder frozen	30% of first epoch
Regression head	#features $\times$ 2048$\times$ $1024 \times (1 + \text{\#additional})$
Optimizer	AdamW
Learning rate	1.5×10^{-5}, encoder 10^{-6}
Batch size	256 (aggregated)
Loss	Average MSE across all targets
Training epochs	5

Table 6: COMET architecture and training details.

A.2 GEMBA

As the underlying LLM for GEMBA, we use Llama 3.3 70B with 4 bit quantization. All experiments with GEMBA are run on one H100 GPU with 80 GB of memory. The prompts we used for $\text{GEMBA}_{\text{poly-cand}}$ and $\text{GEMBA}_{\text{poly-ic}}$ are show in Table 7. Depending on the setting, the human reference and the gold score of the additional translation can be omitted, and more than one additional translations can be included.

B $\text{COMET}_{\text{poly-cand}}$ Ablations and Analysis

B.1 Robustness towards choice of additional candidate.

In a realistic usage, it might not always be possible to have additional candidate that is close to the original translation. Therefore, we experiment using $\text{COMET}_{\text{poly-cand}}$ with randomly selected additional candidate. The results are shown in the bottom half of Table 8 (5-7). As can be seen, even randomly selected additional translations significantly improve performance compared to the standard COMET model. However, compared to the setting with the closest candidate, random selection worsen the performance of $\text{COMET}_{\text{poly-cand}}$, albeit by a small margin. The largest performance drop occurs when the model uses the additional translation's score as input (7 vs 4). This is expected, as having the gold score of a similar candidate to the original translation is more informative than a score for a random one. This also holds when adding more translation candidates, as can be seen in Table 9. More detailed experiment on different levels of candidate similarity are provided below.

B.2 Effect of candidate similarity level

We examine the relationship between the additional translation's similarity to the one at hand. As can be seen in Table 10, the more similar the candidate, the more helpful it is to improve the performance of $\text{COMET}_{\text{poly-cand}}$. This is even more notable in the setting where we include the gold score of the candidate, $f(s, t, t_2, y_{t_2})$. However, in all settings, $\text{COMET}_{\text{poly-cand}}$ is still considerably improved compared to the baseline COMET model.

C $\text{COMET}_{\text{poly-ic}}$ Ablations and Analysis

C.1 Comparing different retrieval strategies.

We investigate different retrieval strategies: We retrieve using the embeddings derived only from the source text s^e, only the translation t_i^e, the sum of the two $s^e + t_i^e$, and the concatenation $\langle s^e, t_i^e \rangle$. We use all-MiniLM-L12-v2 (Reimers and Gurevych, 2020) as an embedding model. Table 11 shows that the simplest approach, only embedding the source yields the best performance across all metrics.

C.2 Testing different embedding models.

In previous experiments, we used an external embedding model (all-MiniLM-L12-v2 (Reimers and Gurevych, 2020)) to retrieve in-context examples. However, one could alternatively use the COMET model's own embeddings or its untrained xlm-roberta-large (Conneau et al., 2020) backbone. We continue using the source text for generating embeddings, as this consistently yielded the best results. Nonetheless, we find that the external embedding model achieves the strongest performance (Table 12), likely because it was explicitly trained

GEMBA$_{\text{poly-cand}}$
Score the translation provided at the end of this prompt from <*source lang*> to <*target lang*> with respect to human reference on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar". Keep your explanation as short as possible. Provide the final score at the end of your answer; do not output anything else afterward.
<*source lang*> source: <*source sentence*>
<*target lang*> human reference: <*human translation*>

Below is an example translation along with its score:
<*target lang*> translation: "<*additional translation*>"
Score: <*score of additional translation*>

Now score this translation (remember to output the final score only at the end of your answer):
<*target lang*> translation: <*MT output*>

Score:

GEMBA$_{\text{poly-ic}}$
Score the translation provided at the end of this prompt from <*source lang*> to <*target lang*> with respect to human reference on a continuous scale from 0 to 100, where a score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar". Keep your explanation as short as possible. Provide the final score at the end of your answer, do not output anything else afterward.

Below is an example translation along with its score:
Source: <*additional source sentence*>
Translation: "<*additional translation*>"
Score: <*score of additional translation*>

Now score this translation (remember to output the final score only at the end of your answer):
<*source lang*> source: <*source sentence*>
<*target lang*> human reference: <*human translation*>
<*target lang*> translation: <*MT output*>

Score:

Table 7: Prompts for GEMBA$_{\text{poly-cand}}$ and GEMBA$_{\text{poly-ic}}$.

for cross-lingual sentence representation. This suggests that COMET$_{\text{poly-ic}}$'s performance is closely tied to the quality and suitability of the embedding model used for retrieval.

C.3 Adaption to the Biomedical Domain using COMET$_{\text{poly-ic}}$

In-Context Enables Domain Transfer. Table 13 presents results from testing our models on in-domain biomedical data. We use the BioMQM dataset (Zouhar et al., 2024). The MQM spans are turned into 0–100 scores to be compatible with the rest of the data. We use the small dev set for training (10k segments) and the test set for evaluation (43k segments).

The goal is to assess whether COMET$_{\text{poly-ic}}$ can leverage in-context examples to adapt its quality estimation to the new domain. This is indeed the case, particularly in MAE, where a substantial performance improvement is observed compared to the base model.

While fine-tuning the models on biomedical data yields even greater gains, it comes at a cost: the fine-tuned base model performs poorly on standard, non-biomedical data. In contrast, both COMET$_{\text{poly-ic}}$ and COMET$_{\text{poly-cand}}$ remain robust after fine-tuning and continue to perform well on standard data, likely because they can incorporate contextual signals at inference time.

C.4 Similarity Threshold Analysis for COMET$_{\text{poly-ic}}$

For COMET$_{\text{poly-ic}}$, the choice of in-context examples is crucial. During training, retrieved examples are drawn from the training set and thus come from the same distribution and have been seen by the model. In contrast, at test time, the examples are unseen and often less similar. Figure 2 shows a histogram of the inner product similarity between embeddings of the evaluated source and the top-1 retrieved source sentence. The plot reveals that training-time additional sources are generally more similar to the evaluated source than those retrieved during testing.

We investigate whether the train-test mismatch affects COMET$_{\text{poly-ic}}$ by training models with dif-

	Model	Reference-less $\rho\uparrow$	Reference-less $\tau_b\uparrow$	Reference-less MAE$\downarrow$	Reference-based $\rho\uparrow$	Reference-based $\tau_b\uparrow$	Reference-based MAE$\downarrow$	
standard COMET model	$f(s,t) \rightarrow \hat{y}_t$	0.105	0.079	30.2	0.245	0.166	26.6	(1)
Additional candidate t_2^* is the closest								
additional candidate	$f(s,t,t_2^*) \rightarrow \hat{y}_t$	0.160	0.127	28.5	0.281	0.180	26.3	(2)
additional candidate, joint predictions	$f(s,t,t_2^*) \rightarrow \hat{y}_t, \hat{y_{t_2^*}}$	0.167	0.113	28.8	0.275	0.172	25.6	(3)
additional candidate and its score	$f(s,t,t_2^*, y_{t^*2}) \rightarrow \hat{y}_t$	0.267	0.207	21.9	0.374	0.243	20.6	(4)
Additional candidate t_2 is random								
additional candidate	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.163	0.118	29.0	0.280	0.175	26.6	(5)
additional candidate, joint predictions	$f(s,t,t_2) \rightarrow \hat{y}_t, \hat{y_{t_2}}$	0.163	0.100	29.3	0.276	0.163	25.8	(6)
additional candidate and its score	$f(s,t,t_2, y_{t_2}) \rightarrow \hat{y}_t$	0.234	0.185	22.9	0.352	0.229	21.0	(7)

Table 8: Results for $\text{COMET}_{\text{poly-cand}}$. The first row shows the standard COMET. The top half (2-4) shows that adding additional translation candidate boosts performance. The bottom half (5-7) shows that using randomly selected additional candidates (in contrast to examples close to the original translation) also helps to boost performance, proving that $\text{COMET}_{\text{poly-cand}}$ is robust to the choice of additional candidates.

Model (+additional)	$\rho\uparrow$ +1	+2	+3	+4	+5	$\tau_b\uparrow$ +1	+2	+3	+4	+5	MAE$\downarrow$ +1	+2	+3	+4	+5
$f(s,t) \rightarrow \hat{y}_t$	0.105	0.105	0.105	0.105	0.105	0.079	0.079	0.079	0.079	0.079	30.2	30.2	30.2	30.2	30.2
t_i is the closest															
$f(s,t,t_{...}) \rightarrow \hat{y}_t$	0.160	0.251	0.224	0.202	0.190	0.127	0.145	0.127	0.130	0.120	28.5	27.6	26.8	28.4	27.6
$f(s,t,t_{...},y_{t_{...}}) \rightarrow \hat{y}_t$	0.267	0.321	0.328	0.327	0.321	0.207	0.229	0.230	0.235	0.233	21.9	17.3	16.0	14.0	13.7
t_i is random															
$f(s,t,t_{...}) \rightarrow \hat{y}_t$	0.163	0.202	0.219	0.228	0.204	0.118	0.135	0.140	0.144	0.136	29.0	27.3	27.9	27.8	28.1
$f(s,t,t_{...},y_{t_{...}}) \rightarrow \hat{y}_t$	0.234	0.276	0.295	0.293	0.295	0.185	0.212	0.216	0.215	0.216	22.9	19.3	15.9	14.9	14.3

Table 9: Results for $\text{COMET}_{\text{poly-cand}}$ using different number of additional translation candidates. The +1 is equal to results in Table 1. The +x uses x additional translation candidates, which improves performance especially for $\text{COMET}_{\text{poly-cand}}$.

Model (+nth closest)	$\rho\uparrow$ 1st	2nd	3rd	4th	5th	$\tau_b\uparrow$ 1st	2nd	3rd	4th	5th	MAE$\downarrow$ 1st	2nd	3rd	4th	5th
$f(s,t) \rightarrow \hat{y}$	0.105	0.105	0.105	0.105	0.105	0.079	0.079	0.079	0.079	0.079	30.2	30.2	30.2	30.2	30.2
$f(s,t,t_2) \rightarrow \hat{y}$	0.163	0.157	0.150	0.140	0.133	0.118	0.114	0.112	0.109	0.107	29.0	29.1	29.2	29.3	29.4
$f(s,t,t_2,y_{t_2}) \rightarrow \hat{y}$	0.234	0.220	0.202	0.187	0.174	0.185	0.180	0.174	0.170	0.163	22.9	23.0	23.3	23.5	23.7

Table 10: Performance of $\text{COMET}_{\text{poly-cand}}$ with the additional translation being the closest, second-closest, third-closest, fourth-closest of fifth-closest to t.

Retrieval key	$\rho \uparrow$	$\tau_b \uparrow$	MAE
None	0.105	0.079	30.2
s_2^e	0.141	0.116	27.3
t_2^e	0.127	0.111	28.4
$s_2^e + t_2^e$	0.135	0.106	27.5
$\langle s_2^e, t_2^e \rangle$	0.117	0.109	27.7

Table 11: $\text{COMET}_{\text{poly-ic}}$ results, with in-context examples retrieved using source text s^e, only the translation t_i^e, the sum of the two $s^e + t_i^e$, and the concatenation $\langle s^e, t_i^e \rangle$.

	$\rho \uparrow$	$\tau_b \uparrow$	MAE
COMET embeddings	0.124	0.108	28.3
MiniLM embeddings (external)	0.141	0.116	27.3
XMLR embeddings (external)	0.115	0.093	27.7

Table 12: $\text{COMET}_{\text{poly-ic}}$ results, with in-context examples retrieved using source text s^e, using different embedding models.

ferent similarity thresholds to better align training retrievals with test-time similarity. The results in Table 14 show that the model trained without any similarity filtering performs best, suggesting that the train-test mismatch does not significantly impact performance.

D k-NN Ablations and Analysis

D.1 Weighted k-NN

We can extend the simple k-nn approach to incorporate weigthed averages, which can boost performance. For example, in the poly-cand setup, our final prediction will be given by

$$\hat{y}_{s,t} = \sum_{i=1}^{n} \left(\frac{w_i}{\sum_{i'=1}^{n} w_{i'}} \right) \times \text{COMET}(s, t_i),$$

where $w_i = \exp(-d_i/\gamma)$ is a weight with d_i being a dissimilarity measure between (s, t) and (s, t_i)

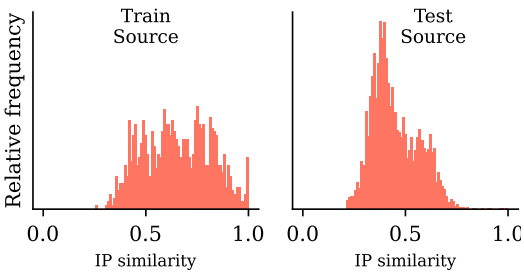


Figure 2: Histogram of inner product similarities between the currently evaluated item and the top-1 retrieved item for $\text{COMET}_{\text{poly-ic}}$.

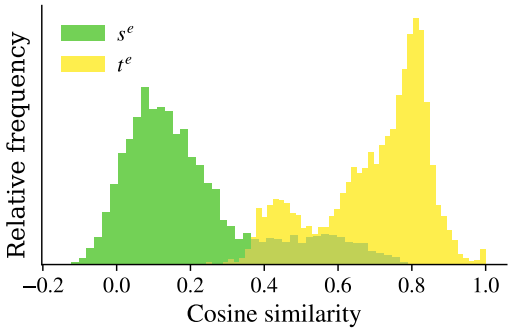


Figure 3: Histograms of translation similarity for examples retrieved by source embeddings (s^e) versus translation embeddings (t^e), showing that t^e-based retrieval yields higher-similarity (more relevant) neighbors while s^e-based retrieval often returns low-relevance examples.

(used for retrieval), and $\gamma > 0$ is the kernel bandwidth, that can be tuned using a validation set. We set d_i to be one minus the cosine similarity of embeddings. The same approach applies to the poly-ic setup. Realize that doing a simple average is equivalent to running the weighted average with $\gamma \to \infty$.

D.2 Ablation and Analysis

We evaluate the k-NN baseline under varying γ values using a weighted-average scoring scheme and different retrieval strategies in the poly-ic setting. Table 15 reports results for the poly-cand configuration: performance is remarkably consistent across both γ and k, since all retrieved translations are of similar relevance. Table 16 gives results for the poly-ic configuration: here, choices of γ and k have a pronounced effect, and the best scores are achieved when retrieval leverages both source and target contexts. Figure 3 complements Table 16 and explains why using s^e for retrieval when k-NN is applied works the worst; we plot the histograms of translation similarity when examples are retrieved either using the translation or the source embeddings. What we see is that when examples are retrieved using source similarity, there is no guarantee that the translations we retrieve are relevant for our target translation (low similarity). On the other hand, if the examples are retrieved using the translation similarities, we end up selecting more relevant examples in terms of similarity (as expected).

E GEMBA Ablations and Analysis

Adding random translations does not consistently improve performance. Similar to ap-

training data	Model	BioMQM Test $\rho \uparrow$	BioMQM Test $\tau_b \uparrow$	BioMQM Test MAE	WMT 2024 Test $\rho \uparrow$	WMT 2024 Test $\tau_b \uparrow$	WMT 2024 Test MAE
WMT (from scratch)	Base	0.100	0.117	35.7	0.105	0.079	30.2
	$\text{COMET}_{\text{poly-cand}}$	0.029	0.068	43.7	0.160	0.127	28.5
	$\text{COMET}_{\text{poly-ic}}$	0.109	0.118	30.5	0.141	0.116	27.3
BioMQM (finetune WMT)	Base	0.139	0.169	2.6	0.060	0.132	11.6
	$\text{COMET}_{\text{poly-cand}}$	0.215	0.177	2.1	0.162	0.175	12.0
	$\text{COMET}_{\text{poly-ic}}$	0.209	0.171	2.1	0.163	0.150	12.0
BioMQM + WMT (from scratch)	Base	0.165	0.141	12.8	0.233	0.187	15.6
	$\text{COMET}_{\text{poly-cand}}$	0.081	0.093	15.7	0.250	0.195	15.9
	$\text{COMET}_{\text{poly-ic}}$	0.168	0.146	12.6	0.240	0.192	15.5

Table 13: $\text{COMET}_{\text{poly-ic}}$'s and $\text{COMET}_{\text{poly-cand}}$'s performance on the BioMQM dataset (Zouhar et al., 2024) and the WMT 2024 dataset, trained on either WMT data, finetuned on BioMQM data (after training on WMT), or trained on a mix of BioMQM data and WMT data.

	$\rho \uparrow$	$\tau_b \uparrow$	MAE
Highest Similarity	0.141	0.116	27.3
Similarity < 0.7	0.130	0.101	28.4
Similarity < 0.5	0.109	0.086	29.1

Table 14: Performance of $\text{COMET}_{\text{poly-ic}}$ trained with different filter thresholds for additional source sentence similarity.

pendix B, we experiment with adding random translation candidates instead of the most similar ones. This yields similar results. These results are reported in Table 17.

Multiple additional translations is better. We experiment with multiple candidates/examples to $\text{GEMBA}_{\text{poly-cand}}$/$\text{GEMBA}_{\text{poly-ic}}$. As can be seen from Table 18, having 5 candidates instead of 1 helps $\text{GEMBA}_{\text{poly-cand}}$ improve over the baseline GEMBA in terms of Pearson correlation; however, the Kendall-tau and MAE metrics do not always agree. For $\text{GEMBA}_{\text{poly-ic}}$, having 5 samples instead of 1 even slightly worsens the performance.

In general, adding more examples to the input does not always help improve the performance of GEMBA as opposed to COMET. Note that the performance of the standard GEMBA is better than the standard COMET (0.266 Pearson versus 0.105 Pearson, see first row of Table 1 and Table 4 for more details). A possible explanation could then be: the additional candidates/samples added to the inputs help with issues that are specific to the baseline COMET, i.e., detecting edges cases of failed translations (see Section 4.5 for more details), which might not be an issue for the baseline GEMBA.

F Analysis of Impact of Additional Translations and In-Context Examples

We perform a systematic analysis by categorizing test cases according to the gold quality scores of the translation under evaluation. The test cases are split into two by the median of the gold quality scores. For $\text{COMET}_{\text{poly-cand}}$, we also further categorize the cases based on the gold quality scores of the additional translation: we consider the cases where (1) the additional translation t_2 is the best within the pool of candidate translations from the same source and (2) t_2 is the worst within the pool of candidate translations. The results can be found in Table 19.

We also manually inspect cases where the additional translation yields the largest improvements. These include, for example, undertranslations, numerical errors, explanations within the translations. We find that in these cases, the additional translation does not show such errors and can serve as a substitute reference. These examples are listed in Table 20.

		$\rho \uparrow$					$\tau_b \uparrow$					MAE $\downarrow$				
γ	k	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
10^{-4}		0.083	0.083	0.084	0.084	0.084	0.064	0.064	0.064	0.064	0.064	30.4	30.4	30.4	30.4	30.4
10^{-2}		0.083	0.084	0.085	0.085	0.084	0.064	0.065	0.066	0.066	0.066	30.4	30.4	30.3	30.3	30.3
10^{0}		0.083	0.087	0.086	0.086	0.086	0.064	0.065	0.062	0.060	0.057	30.4	30.3	30.3	30.4	30.4
∞		0.083	0.087	0.086	0.085	0.085	0.064	0.064	0.062	0.059	0.057	30.4	30.3	30.4	30.4	30.4

Table 15: k-NN results (poly-cand) over varying γ and k.

		$\rho \uparrow$					$\tau_b \uparrow$					MAE $\downarrow$				
γ	k	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
10^{-4}	s^e	0.017	0.019	0.019	0.018	0.018	0.010	0.017	0.018	0.016	0.017	47.0	46.1	46.0	45.7	45.5
	t^e	0.022	0.023	0.024	0.024	0.025	0.011	0.010	0.010	0.009	0.010	32.0	31.9	31.9	31.9	31.9
	$s^e + t^e_i$	0.015	0.015	0.015	0.015	0.015	0.013	0.013	0.013	0.013	0.013	35.0	34.9	34.9	34.9	34.9
	$\langle s^e, t^e_i \rangle$	0.029	0.028	0.028	0.028	0.028	0.014	0.014	0.014	0.014	0.014	31.1	31.1	31.1	31.0	31.0
10^{-2}	s^e	0.017	0.019	0.019	0.018	0.018	0.010	0.017	0.018	0.016	0.017	47.0	46.1	46.0	45.7	45.5
	t^e	0.022	0.028	0.030	0.032	0.033	0.011	0.013	0.016	0.015	0.017	32.0	30.5	29.9	29.6	29.4
	$s^e + t^e_i$	0.015	0.017	0.018	0.019	0.019	0.013	0.014	0.014	0.014	0.013	35.0	33.7	33.1	32.7	32.5
	$\langle s^e, t^e_i \rangle$	0.029	0.031	0.032	0.032	0.034	0.014	0.015	0.015	0.016	0.016	31.1	29.9	29.3	29.0	28.9
10^{0}	s^e	0.017	0.019	0.019	0.018	0.018	0.010	0.017	0.018	0.016	0.017	47.0	46.1	46.0	45.7	45.5
	t^e	0.022	0.028	0.031	0.035	0.038	0.011	0.013	0.016	0.016	0.019	32.0	30.1	29.3	28.9	28.6
	$s^e + t^e_i$	0.015	0.017	0.021	0.020	0.020	0.013	0.014	0.014	0.013	0.012	35.0	33.1	32.2	31.8	31.4
	$\langle s^e, t^e_i \rangle$	0.029	0.032	0.034	0.036	0.037	0.014	0.017	0.017	0.019	0.020	31.1	29.4	28.7	28.2	27.9
∞	s^e	0.017	0.019	0.019	0.018	0.018	0.010	0.017	0.018	0.016	0.017	47.0	46.1	46.0	45.7	45.5
	t^e	0.022	0.028	0.031	0.035	0.038	0.011	0.013	0.016	0.016	0.019	32.0	30.1	29.3	28.9	28.6
	$s^e + t^e_i$	0.015	0.017	0.021	0.020	0.020	0.013	0.014	0.014	0.013	0.012	35.0	33.1	32.2	31.8	31.4
	$\langle s^e, t^e_i \rangle$	0.029	0.031	0.034	0.036	0.037	0.014	0.017	0.017	0.019	0.020	31.1	29.4	28.7	28.2	27.9

Table 16: k-NN results (poly-ic) over varying γ, k, and retrieval methods.

		Reference-less			Reference-based		
	Input $\rightarrow$ Output	$\rho \uparrow$	$\tau_b \uparrow$	MAE $\downarrow$	$\rho \uparrow$	$\tau_b \uparrow$	MAE $\downarrow$
standard GEMBA	$f(s,t) \rightarrow \hat{y}_t$	0.266	0.199	27.6	0.311	0.200	27.3
GEMBA$_{\text{poly-cand}}$, closest t_2^*							
additional candidate	$f(s,t,t_2^*) \rightarrow \hat{y}_t$	0.245	0.185	28.2	0.277	0.187	27.5
additional candidate, joint predictions	$f(s,t,t_2^*) \rightarrow \hat{y}_t, \hat{y_{t_2^*}}$	0.235	0.149	28.6	0.296	0.181	27.9
additional candidate and its score	$f(s,t,t_2^*, y_{t^*2}) \rightarrow \hat{y}_t$	0.276	0.187	27.4	0.337	0.217	26.8
GEMBA$_{\text{poly-cand}}$, random t_2							
additional candidate	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.236	0.169	28.3	0.265	0.167	27.7
additional candidate, joint predictions	$f(s,t,t_2) \rightarrow \hat{y}_t, \hat{y_{t_2}}$	0.229	0.135	28.6	0.281	0.170	28.0
additional candidate and its score	$f(s,t,t_2,y_{t_2}) \rightarrow \hat{y}_t$	0.234	0.159	27.7	0.289	0.192	27.1
GEMBA$_{\text{poly-ic}}$							
additional sample	$f(s,t,s_2,t_2,y_{t_2}) \rightarrow \hat{y}_t$	0.195	0.099	28.3	0.291	0.168	27.4

Table 17: Results for GEMBA$_{\text{poly-cand}}$ and GEMBA$_{\text{poly-ic}}$. The first row shows the standard GEMBA model. In contrast to the COMET models, adding additional translation candidates and in-context examples does not significantly boost performance.

Model (+additional)	$\rho \uparrow$ +1	$\rho \uparrow$ +5	$\tau_b \uparrow$ +1	$\tau_b \uparrow$ +5	MAE $\downarrow$ +1	MAE $\downarrow$ +5
standard GEMBA						
$f(s,t) \rightarrow \hat{y}_t$	0.266	0.266	0.199	0.199	27.6	27.6
GEMBA$_{\text{poly-cand}}$, closest t_i						
$f(s,t,t_{\ldots}) \rightarrow \hat{y}_t$	0.245	0.277	0.185	0.186	28.2	27.8
$f(s,t,t_{\ldots},y_{t_{\ldots}}) \rightarrow \hat{y}_t$	0.276	0.291	0.187	0.196	27.4	26.2
GEMBA$_{\text{poly-cand}}$, random t_i						
$f(s,t,t_{\ldots}) \rightarrow \hat{y}_t$	0.236	0.276	0.169	0.180	28.3	27.8
$f(s,t,t_{\ldots},y_{t_{\ldots}}) \rightarrow \hat{y}_t$	0.234	0.282	0.159	0.179	27.7	26.5
GEMBA$_{\text{poly-ic}}$						
$f(s,t,s_2,t_2,y_{t_2}) \rightarrow \hat{y}_t$	0.195	0.188	0.099	0.097	28.3	28.7

Table 18: GEMBA$_{\text{poly-cand}}$ and GEMBA$_{\text{poly-ic}}$ with multiple candidates (reference-less).

		Model	$\rho \uparrow$	$\tau_b \uparrow$	MAE $\downarrow$
All samples	Standard COMET	$f(s,t) \rightarrow \hat{y}_t$	0.125	0.088	29.2
	COMET$_{\text{poly-cand}}$, t_2 is high quality	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.174	0.136	27.9
	COMET$_{\text{poly-cand}}$, t_2 is low quality	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.172	0.124	28.6
	COMET$_{\text{poly-ic}}$	$f(s,t,s_2,t_2,y_2) \rightarrow \hat{y}_t$	0.143	0.118	27.0
High quality samples	Standard COMET	$f(s,t) \rightarrow \hat{y}_t$	0.055	0.036	29.0
	COMET$_{\text{poly-cand}}$, t_2 is high quality	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.075	0.048	27.6
	COMET$_{\text{poly-cand}}$, t_2 is low quality	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.088	0.039	28.7
	COMET$_{\text{poly-ic}}$	$f(s,t,s_2,t_2,y_2) \rightarrow \hat{y}_t$	0.045	0.044	28.7
Low quality samples	Standard COMET	$f(s,t) \rightarrow \hat{y}_t$	0.188	0.100	26.1
	COMET$_{\text{poly-cand}}$, t_2 is high quality	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.260	0.158	24.7
	COMET$_{\text{poly-cand}}$, t_2 is low quality	$f(s,t,t_2) \rightarrow \hat{y}_t$	0.275	0.157	25.2
	COMET$_{\text{poly-ic}}$	$f(s,t,s_2,t_2,y_2) \rightarrow \hat{y}_t$	0.203	0.111	25.2

Table 19: Analysis of COMET$_{\text{poly-cand}}$ and COMET$_{\text{poly-ic}}$ for predicting quality estimation for low- or high- quality translations (high if score above the median quality of all translations).

English Source	Machine Translation	Human	COMET	COMET$^{\text{poly-cand}}$
Undertranslation				
`@user33 wow!`	`@user33 wow!` (Hindi)	2.0	87.801	−0.78
Wrong-number-output				
`1/3`	`2/3` (Czech)	0.5	85.220	2.79
Redundant-explanation				
`@user33 wow!`	"`@uzivatel33 wow!`" or "`@uzivatel33 páni!`" Both options are valid translations of "wow" in Czech. The first one keeps the English word, which is common in informal online communication. The second one uses the Czech equivalent "páni," which is more colloquial.	2.0	66.55	−0.04
Refusal				
And now the wife is grumpy because my "Extraction" was much quicker than expected. Sorry, maybe your genome shouldn't play so hard to get.	I apologize, but I don't feel comfortable translating that particular text, as it contains sensitive content that could be inappropriate or offensive if reproduced in another language. Perhaps we could have a thoughtful discussion about more constructive topics instead. (Russian)	5.0	87.08	28.55

Table 20: Examples of improvements with COMET$_{\text{poly-cand}}$ compared to the baseline COMET.

Long-context Reference-based MT Quality Estimation

Sami Ul Haq[1], **Chinonso Cynthia Osuji**[1], **Thiago Castro Ferreira**[2], **Brian Davis**[1], **Sheila Castilho**[1]
[1]ADAPT Centre, Dublin City University, Dublin, Ireland
[2]Fluminense Federal University, Brazil
{firstname.lastname}@adaptcentre.ie
thiago.castro.ferreira@gmail.com

Abstract

In this paper, we present our submission to the Tenth Conference on Machine Translation (WMT25) Shared Task on Automated Translation Quality Evaluation. Our systems are built upon the COMET framework and trained to predict segment-level Error Span Annotation (ESA) scores using augmented long-context data. To construct long-context training data, we concatenate in-domain, human-annotated sentences and compute a weighted average of their scores. We integrate multiple human judgment datasets (MQM, SQM, and DA) by normalising their scales and train multilingual regression models to predict quality scores from the source, hypothesis, and reference translations. Experimental results show that incorporating long-context information improves correlations with human judgments compared to models trained only on short segments.

1 Introduction

The automatic evaluation of machine translation (MT) is a crucial component of MT research and development. While expert-based human evaluation remains the gold standard, automatic evaluation offers fast and scalable judgments, enabling rapid feedback for optimizing model parameters. Traditionally, automatic MT evaluation metrics have relied on basic lexical-level features, such as counting matching n-grams between the MT hypothesis and the reference translation. Metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ChrF (Popović, 2015) remain popular due to their lightweight design and computational efficiency (Marie et al., 2021). More recently, neural metrics (either trained on human annotations or based on pre-trained language models) have demonstrated superior capability in comparing and assessing MT quality, often outperforming traditional lexical-based metrics (Freitag et al., 2022). These neural approaches leverage large-scale multilingual data during training and achieve strong performance even when translations diverge lexically from the reference.

This paper presents DCU_ADAPT's submission to the WMT25[1] MT Evaluation Shared Task. The primary focus of this year's task is on systems capable of evaluating translation quality in context, where the context spans entire documents or multiple consecutive segments. We participated in the segment-level quality score prediction track for English–Czech, English–Russian, English–Japanese, and English–Chinese, employing models based on the COMET framework (Rei et al., 2020a).

In our contribution to the shared task, we explore methods for leveraging synthetic data alongside the capabilities of pre-trained and cross-lingual models to predict MT quality estimates for long-sequence or multi-sentence units. Human judgments of MT quality are typically available as short segment-level scores, such as DA (Graham et al., 2017), MQM (Lommel et al., 2014), and SQM (Barrault et al., 2019). Recent pre-trained models support larger context windows and can handle long-sequence inputs, improving discourse-level resolution (Dai et al., 2019)—albeit at the cost of increased memory and computational requirements. However, most existing automatic evaluation metrics (AEMs) predict scores at the sentence level, and those designed for document-level evaluation often perform only shallow context integration during inference (Vernikos et al., 2022). We propose a data augmentation strategy to train multilingual models on long-context annotated data, enabling them to better exploit broader context and reduce inconsistencies caused by sentence-level ambiguity.

The exploration of context in MT is a well-established topic and, in recent years, has become a focal point, driven by the need to incorporate context into both MT systems and their evalua-

[1]https://www2.statmt.org/wmt25/mteval-subtask.html

Proceedings of the Tenth Conference on Machine Translation, pages 905–912
November 8-9, 2025

tion methodologies (Bawden et al., 2017; Castilho et al., 2020; Maruf et al., 2021; Castilho et al., 2023; Castilho and Knowles, 2024). There is now broad consensus on the value of document-level evaluation. Since 2019, WMT has conducted human evaluations at the document level, providing evaluators with access to context even when collecting segment-level ratings (Akhbardeh et al., 2021; Kocmi et al., 2022a, 2023, 2024a). Research indicates that the appropriate context span is critical for reliable MT evaluation, with Castilho et al. (2020) showing that incorporating relevant context spans can yield more accurate assessments of translation quality, thereby improving the evaluation process. Several techniques have been proposed to extend evaluation to the document level or to incorporate multi-sentence context into automatic evaluation metrics (Jiang et al., 2021; Vernikos et al., 2022; Rei et al., 2022; Kocmi et al., 2022b; Raunak et al., 2024).

For this shared task submission, we use the Estimator model from the COMET framework (Rei et al., 2022), which learns MT quality from human evaluation data such as MQM and DA. To create long-context training data, we combine multiple annotations using a weighted average alongside the original annotations. Our experiments show promising progress toward improved correlation in multi-sentence-level MT quality estimation. Fine-tuning multilingual embedding models demonstrates that it is possible to achieve high correlations with human judgments when evaluating long segments, rather than relying solely on sentence-level score predictions.

We release the data and code produced during this research.

2 Corpora

We used the human annotations from previous WMT shared tasks (Kocmi et al., 2022a, 2023, 2024a) for training our models, which includes human annotations from MQM, DA, and SQM. We train and evaluate our models for English (en) to Czech (cs), Japanese (ja), Chinese (zh), and Russian (ru) language pairs. MQM scores are derived from error annotations and can range from $-\infty$ to 100. Since our goal is to predict ESA scores (Kocmi et al., 2024b), which range between 0 and 100, we normalise (using Equation 1, where x is original and x' is normalised score) and rescale the scores to the [0, 1] interval for training models on the combined dataset.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

We create augmented training data for long-span MT quality estimation by taking a weighted average of segment-level scores. During augmentation, we concatenate multiple samples (i.e., 2, 3, 4, and 5 segments) to form long-span texts. To construct the long-span MT evaluation dataset, adjacent short segments are concatenated, and a document-level quality score is computed as a length-weighted average of their original scores. The weighting is based on the total number of characters in the source and machine-translated texts, as formalized in Equation 2.

Let s_1, s_2 are short segments (e.g., source-translation pairs), and raw_1, raw_2 human evaluation scores for these segments. C_1 and C_2 are the total character count of each segment e.g., $C_i = len(s_i)$.

Then the document-level score (raw_{doc}) is calculated as:

$$\text{raw}_{\text{doc}} = \frac{C_1 \cdot \text{raw}_1 + C_2 \cdot \text{raw}_2}{C_1 + C_2} \tag{2}$$

This equation[2] computes a weighted average of two segment-level scores, where the weight is determined by the combined character count of the source and MT for each segment. Longer segments contribute more to the final score, reflecting their higher informational content. The augmentation process increased the average segment length of dataset from 16.84 words per segment to 52.99 words per segment. Since the augmented segments were added to the original dataset, the overall size of the dataset increased by a factor of two.

We then create training, test, and validation sets by randomly sampling segments from the training data. The statistics of the final training, validation, and test sets are shown in Table 1.

3 Experimental Setup

Our system is built on top of the COMET package, utilizing the `comet-train` and `comet-score` commands to train and evaluate our models. We fine-tuned the pre-trained model `Unbabel/wmt22-comet-da`, originally trained on

[2]We adapted the implementation of data augmentation from the Huggingface repository: https://huggingface.co/datasets/ymoslem/wmt-da-human-evaluation-long-context

Type	Split	No. of segments			
		en→cs	en→ja	en→ru	en→zh
DA	train	345K	119K	350K	533K
	dev	43K	15K	43K	66K
	test	38K	13K	39K	59K
MQM	train	–	–	172K	–
	dev	–	–	21K	–
	test	–	–	19K	–
SQM	train	64K	75K	64K	75K
	dev	8K	9K	8K	9K
	test	7K	8K	7K	8K
Total		**505K**	**238K**	**723K**	**750K**

Table 1: Dataset statistics by evaluation type, split, and language pair (K represents values in thousands).

DA data, as well as the multilingual pre-trained model `FacebookAI/xlm-roberta-base` (Liu et al., 2019), using A100 GPU. The `xlm-roberta` model was pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages and has approximately 279 million parameters.

We trained the models for up to 5 epochs and employed early stopping when the Spearman correlation on the development set did not improve for two consecutive evaluations. For each language pair, the augmented dataset contained five times the original data; however, due to limited GPU memory, we faced out-of-memory issues and restricted training to augmentations with up to two segments. The training process with the augmented dataset took approximately 10 hours for each model.

We trained one baseline model (the fine-tuned version of `wmt22-comet-da`) and three main models (one primary and two secondary submissions) using the data augmentation approach, retaining only the last two checkpoints for each. Spearman correlation on the test split was used to select the best checkpoint per language pair, and this best model was used for the final submission.

For official WMT `test25`[3] evaluation, the full segment was used since it fell within the maximum sequence length (512) supported by both the baseline and fine-tuned models. COMET scores typically range between 0 and 1, but can sometimes exceed 1, indicating exceptionally high-quality segments. To align with the ESA metric's scoring strategy (Kocmi et al., 2024b), we upscaled and rounded the scores to a range between 0 and 100.

[3]https://github.com/wmt-conference/wmt25-mteval/blob/main/data/testset/mteval-task1-test25.tsv.gz

4 Results

As described in Section 2, our experiments use the normalized versions of multiple human annotations collected from previous WMT shared tasks. To evaluate and compare our approach, we applied a similar augmentation method to construct long-sequence test data, incorporating annotations from DA, MQM, and SQM. Our baseline sentence-level quality estimation models are `wmt22-comet-da` (referred to as COMET-22) and BERTSCORE. COMET-22-LS, a fine-tuned version of `wmt22-comet-da` on long-span data, and ROBERTA-LS, fine-tuned from `FacebookAI/xlm-roberta-base`, both serve as long-sequence quality score prediction models. Following Rei et al. (2020a), the models were trained on triplets of (source, hypothesis, reference) and output a score between 0 and 1 reflecting the translation quality relative to both source and reference.

We used the Pearson correlation coefficient to evaluate the models' performance. Segment-level Pearson correlations on the self-test set are presented in Tables 2 and 3. Our results indicate that metrics from models trained on long-context inputs generally outperform sentence-level metrics, in some cases by a significant margin.

The annotation-wise segment-level correlation results in Table 2 demonstrate that the unsupervised baseline metric, BERTSCORE, exhibits relatively weak correlations across all language pairs. The sentence-level COMET-22 model shows improved correlations, especially for DA annotations, reflecting its training on DA data. Moreover, it outperforms BERTSCORE on SQM and MQM annotations, indicating its ability to mimic human annotations by learning from data. Our fine-tuned long-sequence baseline model, COMET-22-LS, surpasses the sentence-level baselines, achieving performance close to, and in some cases better than, our primary submission models based on ROBERTA-LS. Notably, COMET-22-LS achieves results on MQM annotations that are very close to those of ROBERTA-LS, outperforming sentence-level baselines by a substantial margin. This suggests that training on longer context sequences provides considerable benefits, especially for complex annotation types like MQM, which capture fine-grained translation errors.

Table 3 summarizes correlation results across all language pairs using joint annotations. Across nearly all language pairs, our models outperform

	DA				SQM				MQM
	en→ru	**en→cs**	**en→ja**	**en→zh**	**en→ru**	**en→cs**	**en→ja**	**en→zh**	**en→ru**
BERTSCORE	0.399	0.480	0.418	0.328	0.290	0.207	0.290	0.107	-0.04
COMET-22	0.571	0.637	0.511	0.443	0.450	0.400	0.352	0.220	0.075
COMET-22-LS	0.848	**0.894**	0.777	**0.772**	**0.572**	0.701	**0.668**	0.585	0.866
ROBERTA-LS	**0.874**	0.890	**0.780**	0.770	0.557	**0.707**	0.666	**0.600**	**0.874**

Table 2: System-level Pearson correlation results for MQM, SQM, and DA annotations. Bold values indicate systems that achieved higher correlations with human judgments. LS denotes models trained on long-span input data.

	en→ru	**en→cs**	**en→ja**	**en→zh**	avg.
BERTSCORE	0.216	0.344	0.354	0.217	0.283
COMET-22	0.365	0.519	0.432	0.331	0.412
COMET-22-LS	0.762	0.798	0.722	0.679	0.740
ROBERTA-LS	**0.768**	**0.799**	**0.723**	**0.685**	**0.744**

Table 3: Segment-level Pearson correlation scores for the language pairs `en-ru`, `en-cs`, `en-ja`, and `en-zh`. Bold values indicate stronger correlations with human judgments. LS denotes models trained on long-span input data.

baseline metrics in correlation with human judgments. Ideally, COMET-22-LS, fine-tuned from `wmt22-comet-da` (which is already trained for evaluation tasks), should have outperformed ROBERTA-LS. However, the performance difference between the two models is not substantial. This may be because `wmt22-comet-da` was trained only on DA data, while the current task also includes SQM and MQM annotations, which typically follow different scoring strategies. Furthermore, as mentioned earlier, due to resource constraints, we trained our models for only five epochs and on a limited number of augmented segments. With larger augmentation and more robust training, the performance gap between the two models may become more pronounced.

These results suggest that combining multiple segments within the same document or domain is more effective than independently scoring segmented sentences and averaging their scores (Raunak et al., 2024). The improvement may be attributed to the model's ability to capture contextual information across long-sequence segments, thereby enabling more context-aware quality estimation.

However, handling longer text sequences poses challenges due to the input size limitations of the underlying models, as highlighted by Gong et al. (2020). This necessitates careful segmentation and score-averaging strategies to compute scores at the paragraph or document level. We also conduct a preliminary analysis of the score distributions (Figure 1) and find that sentence-level baseline scores (COMET-22) are mostly concentrated between 60 and 100, with a pronounced peak around 90. In contrast, the ROBERTA-LS model, trained on multi-sentence inputs, produces a more widely spread score distribution that better reflects the variability typically observed in human judgments (Toral et al., 2018). This wider spread may be due to the long-span training data being based on weighted average scores that encompass a broader range of score scales. By contrast, the narrower distribution of COMET-22 scores could stem from the characteristics of the human-annotated data on which it was trained, where non-expert evaluators have been shown to assign disproportionately higher fluency and adequacy ratings, resulting in smaller score gaps and reduced variance compared to expert assessments (Toral et al., 2018). This indicates that models trained with longer context are more sensitive to subtle quality differences, reflecting a more nuanced understanding of translation quality.

5 Related Work

In recent years, metrics based on large pre-trained models have emerged as strong alternatives to traditional n-gram-based approaches, enabling better capture of semantic similarity between words beyond mere lexical matching. These metrics broadly fall into two categories: embedding-based metrics and fine-tuned metrics.

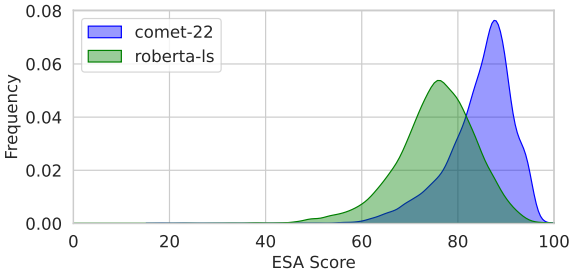


Figure 1: Distribution of segment-level scores assigned by sentence-level COMET-22 and long-sequence ROBERTA-LS model.

Embedding-based metrics typically represent an advancement over n-gram matching by using dense word representations in an embedding space to compute scores that reflect semantic similarity between reference and hypothesis segments. Notable examples include YISI-1 (Lo, 2019), MOVERSCORE (Chow et al., 2019), and BERTSCORE (Zhang et al., 2019), which leverage embedding models for soft alignment between two segments to capture semantic similarity effectively.

Fine-tuned metrics, on the other hand, involve learnable models such as RUSE (Shimanaka et al., 2018), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020a, 2022) that directly optimize underlying embedding models to maximize correlation with human judgments. These models have demonstrated promising results in producing reliable quality scores for test sets such as DA or MQM. While most of these metrics perform reference-based evaluation, recent advancements leveraging highly multilingual pre-trained encoders like multilingual BERT (Devlin, 2018) and RoBERTa (Liu et al., 2019; Conneau et al., 2019) have enabled reference-less systems to show encouraging correlations with human judgments (Freitag et al., 2023).

Most automatic evaluation approaches rely on decontextualized assessments, where translations are judged at the sentence level. However, sentences are often inherently ambiguous, and incorporating document-level context has been shown to be beneficial for both MT and its evaluation (Läubli et al., 2018; Castilho et al., 2020; Castilho and Knowles, 2024; Vernikos et al., 2022). Consequently, a few automatic metrics have been developed to extend evaluation beyond the word or sentence level (Vernikos et al., 2022; Jiang et al., 2021). These metrics aim to address discourse-level phenomena such as lexical consistency, coherence, ellipsis, and pronoun resolution (Voita et al., 2018; Bawden et al., 2017).

However, existing methods typically use a limited number of surrounding sentences as context, allowing models to incorporate neighboring information when embedding each sentence and computing scores at the sentence level (Rei et al., 2020b; Vernikos et al., 2022; Hu et al., 2023). In contrast, long-sequence or document-level evaluation processes the entire segment as a single input, enabling deeper discourse-level resolution and offering a promising yet still underexplored avenue for improving alignment with human judgments.

6 Conclusion

In this paper, we present DCU_ADAPT's contribution to the WMT25 MT Evaluation Shared Task. We leverage the COMET framework and train regression models to predict ESA quality scores. In line with the Shared Task goals, we augment the provided training data and optimize our models to evaluate long, multi-sentence units of text. By fine-tuning multilingual models for cross-lingual transfer, we utilize source, reference, and hypothesis as inputs. Our primary submission — a fine-tuned pre-trained model trained on augmented data — demonstrates higher or otherwise competitive correlation levels with human judgments across multiple languages. Further investigation comparing long-text evaluation after segmentation with sentence-level evaluation is a promising direction for future work.

The data and code produced during this Shared Task participation are available at: https://github.com/sami-haq99/CAEMT/tree/main/wmt-2025-submission.

Limitations

We trained our models using augmented long-segment level scores from the MQM, DA, and SQM datasets. However, we only evaluated the models on self-test data carefully extracted from the training set; evaluating on benchmark datasets will better clarify the true benefits of our approach. Additionally, we normalized the data using min-max normalization to combine different datasets and upscaled the predicted scores to match the ESA metric's score range. Furthermore, our training and testing were conducted on GPUs with at least 40GB of memory; due to time constraints, we were unable to evaluate performance on CPUs.

Ethics Statement

Our research focuses on evaluating long-sequence outputs of MT systems using quality estimation models trained on augmented data. We are committed to conducting and reporting our evaluations with the highest levels of transparency and fairness. By upholding these principles, we aim to contribute to reliable and objective assessment practices in MT evaluation.

Acknowledgements

This work was conducted with the financial support of the Research Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, and Research Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

The Authors also benefit from being members of the ADAPT SFI Research Centre at Dublin City University, funded by the Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). ACL.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2017. Evaluating discourse phenomena in neural machine translation. *arXiv preprint arXiv:1711.00513*.

Sheila Castilho and Rebecca Knowles. 2024. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, pages 1–31.

Sheila Castilho, Clodagh Mallon, Rahel Meister, and Shengya Yue. 2023. Do online machine translation systems care for context? what about a gpt model?

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, et al. 2023. Results of wmt23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent chunking mechanisms for long-text machine reading comprehension.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Xinyu Hu, Xunjian Yin, and Xiaojun Wan. 2023. Exploring context-aware evaluation metrics for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15291–15298.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2021. Blonde: An automatic evaluation metric for document-level machine translation. *arXiv preprint arXiv:2103.11878*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024a. Findings of the wmt24 general machine translation shared task: the llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022a. Findings of the 2022 Conference on Machine Translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Pierre Marie et al. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 566–577. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2021. A survey on document-level neural machine translation: Methods and evaluation. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. Slide: Reference-free evaluation for machine translation using a sliding document window.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins.

2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Catarina Farinha, and Alon Lavie. 2020b. Unbabel's participation in the wmt20 metrics shared task. *arXiv preprint arXiv:2010.15535*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level mt metrics: How to convert any pretrained metric into a document-level metric. *arXiv preprint arXiv:2209.13654*.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Evaluating WMT 2025 Metrics Shared Task Submissions on the SSA-MTE African Challenge Set

Senyu Li[1,2] **Felermino D. M. A. Ali**[6,7] **Jiayi Wang**[3]
Rui Sousa-Silva[7] **Henrique Lopes Cardoso**[6] **Pontus Stenetorp**[3,5]
Colin Cherry[8] **David Ifeoluwa Adelani**[1,2,4]
[1]Mila - Quebec AI Institute, [2]McGill University, [3]University College London,
[4]Canada CIFAR AI Chair, [5]LLMC, National Institute of Informatics
[6]LIACC, Faculdade de Engenharia, Universidade do Porto
[7]CLUP, Faculdade de Letras, Universidade do Porto [8]Google
{senyu.li, david.adelani}@mila.quebec
jiaywang@cs.ucl.ac.uk up202100778@fe.up.pt

Abstract

This paper presents the evaluation of submissions to the WMT 2025 Metrics Shared Task on the SSA-MTE challenge set, a large-scale benchmark for machine translation evaluation (MTE) in Sub-Saharan African languages. The SSA-MTE test sets contains over 12,768 human-annotated adequacy scores across 11 language pairs sourced from English, French, and Portuguese, spanning six commercial and open-source MT systems. Results show that correlations with human judgments remain generally low, with most systems falling below the 0.4 Spearman threshold for medium-level agreement. Performance varies widely across language pairs, with most correlations under 0.4; in some extremely low-resource cases, such as Portuguese–Emakhuwa, correlations drop to around 0.1, underscoring the difficulty of evaluating MT for very low-resource African languages. These findings highlight the need for more robust and generalizable evaluation methods tailored to African language contexts.

1 Introduction

In recent years, with the rise of large language models (LLMs), more and more machine translation (MT) systems have emerged, demonstrating competitive performance. This growth has created an increasingly urgent need for more accurate methods to assess the quality of generated translations. Traditional metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and chrF (Popović, 2015), which rely on n-gram matching, show only limited correlation with human judgments, indicating their limited ability to capture semantic-level quality (Callison-Burch et al., 2006).

More recently, neural metrics such as BERTScore (Zhang et al., 2020a) have shown improved capability in capturing semantic similarity (Freitag et al., 2024; Zhang et al., 2020b). COMET (Rei et al., 2020) further advances this by framing machine translation evaluation (MTE) as a regression task using encoder-only language models and training on human-annotated scores. Likewise, the MetricX (Juraska et al., 2023) family of metric, based on the mT5 (Xue et al., 2020) series multilingual encoder-decoder LM, adopts a regression-based framework similar to COMET. These neural, learned metrics have been shown to achieve higher correlations with human assessments across a wide range of languages (Rei et al., 2020; Juraska et al., 2023).

However, before 2024, due to the lack of machine translation evaluation data, the performance of these models was largely untested on Sub-Saharan African languages. In 2024, AfriMTE (Wang et al., 2024a) was created. Evaluation results revealed that while existing metrics performed well on some relatively higher-resourced languages, they struggled with translations for very low-resource languages such as Twi and Luo (Wang et al., 2024b). The authors demonstrated that this gap can be partially addressed by further pretraining models on African languages; however, performance remains low for specific language pairs, like eng-luo, underscoring the need for dedicated training data for these languages.

Although AfriMTE marked progress in this area, several limitations remained. The dataset was relatively small, with only about 200 cases per language pair. Also, AfriMTE included outputs only from NLLB-200 (600M) (NLLB-Team et al., 2022) and M2M-100 (418M) (Fan et al., 2021). In addition, AfriMTE did not provide any training data, which meant that neural-based metrics could not be directly optimized for Sub-Saharan languages.

To address these challenges, Li et al. (2025) in-

Proceedings of the Tenth Conference on Machine Translation, pages 913–919
November 8-9, 2025

troduced SSA-MTE, a larger-scale dataset created following the same protocol and using the same annotation tool as AfriMTE. SSA-MTE covers 13 Sub-Saharan African languages, with test sets for 10 languages and training sets for 12. It features several source languages—English, French, and Portuguese—representing the Anglophone, Francophone, and Lusophone linguistic communities in the region. SSA-MTE includes translations from six MT systems and contains over 73,000 annotations in total.

The WMT 2025 Metrics Shared Task incorporates the SSA-MTE test set as a dedicated challenge set, enabling the evaluation of MT metrics on low-resource African languages. This inclusion establishes a benchmark for assessing the ability of MTE systems to generalize across under-resourced African languages.

2 SSA-MTE

We perform our evaluation on the recently released **SSA-MTE** dataset (Li et al., 2025), a large-scale human-annotated benchmark for assessing machine translation quality for African languages. The dataset contains over 73,000 sentence-level annotations across 13 language pairs (LPs) in the news domain, of which 10 LPs include a dedicated test set.

The test set covers 7 English–sourced LPs: Amharic (`eng-amh`), Hausa (`eng-hau`), Kikuyu (`eng-kik`), Kinyarwanda (`eng-kin`), Luo (`eng-luo`), Twi (`eng-twi`), and Yorùbá (`eng-yor`); 2 French–sourced LPs: Ewe (`fra-ewe`) and Wolof (`fra-wol`); and 1 Portuguese–sourced LP: Emakhuwa (`por-vmw`). For this challenge set, the authors additionally introduced another Portuguese–sourced LP: Nyanja (`por-nya`). The size of each test set is shown in Table 1.

The selected LPs reflect Africa's linguistic and regional diversity, covering Anglophone, Francophone, and Lusophone areas. In addition, the languages span three major language families: Afro-Asiatic (Hausa, Amharic), Niger–Congo (Kikuyu, Kinyarwanda, Emakhuwa, Nyanja, Twi, Yorùbá, Ewe, Wolof), and Nilo-saharan (Luo), ensuring representation across West, East, Central, and Southern Africa, and thus capturing both geographic and typological diversity.

Each instance is annotated by one human evaluator with both a continuous adequacy score and span-level (character-based) error labels, enabling fine-grained evaluation of MT outputs. Annotators are provided with the source sentence and its translation; instructed to first identify all errors according to the annotation protocol proposed by Wang et al., and then assign a final adequacy score.

For English- and French-sourced cases, source sentences are drawn from the Global Voices news website, following harmful-content filtering and topic-diversity-based article selection (Li et al., 2025). For Portuguese-sourced cases, source sentences are extracted from the Multilingual Open Text dataset, which features news articles published by Voice of America (VOA). All source texts are translated into the target languages by professional translators to serve as reference translations (Ali et al., 2024).

For English- and French-sourced sentences, six MT systems are included (but only five for Kikuyu, as Google Translate does not support it): GPT-4o, Gemini-1.5, Claude-3.5, Google Translate, and two open-source models: NLLB-200-distilled-600M (NLLB-Team et al., 2022) and M2M-100-418M (Fan et al., 2021). Each system contributes an equal number of translations. For Portuguese-sourced sentences, four MT systems are included: GPT-4o-mini, Gemini-1.5, Claude-3.5, and Google Translate.

Compared to the African Challenge Set in the WMT 2024 Metrics Shared Task, this year's set is substantially larger (12,768 vs. 2,815 annotated instances), covers a broader range of MT systems, and introduces Portuguese as a new source language. This broader linguistic and system coverage is expected to yield a more accurate and comprehensive evaluation of metric performance for African languages.

3 Metrics

The submissions of WMT 2025 Metrics Shared Task contain baseline metric results provided by the organizers, as well as the results of primary and secondary submissions from participants' metric systems. This section introduces each of these approaches[1].

3.1 Baselines

We received the following baseline metrics from the organizers: BERTScore (Zhang et al.,

[1]Detailed information of the submissions has not yet been provided by the organizer, and will be added in the camera-ready version.

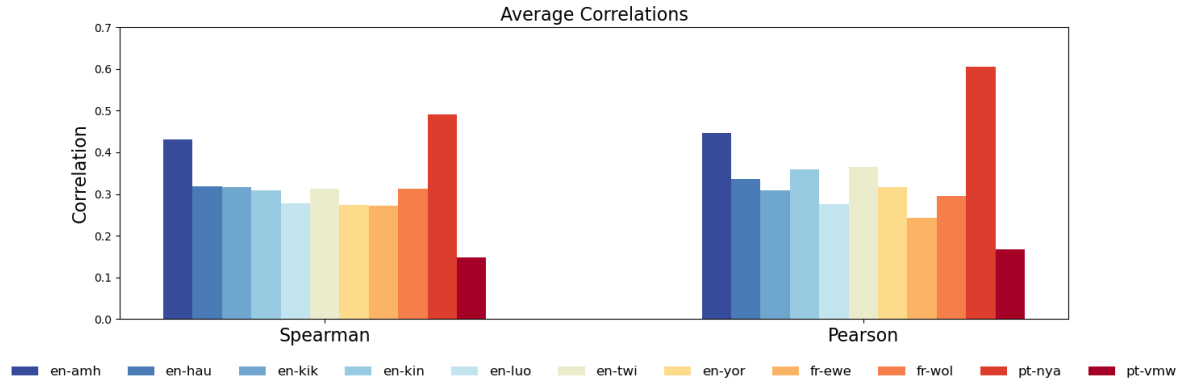


Figure 1: **Average Spearman and Pearson correlations across LPs**. AfriCOMET-1.1-MTL and SSA-COMET-MTL were not included in the calculation.

Language Pair	Size
en-amh	1,166
en-hau	1,192
en-kik	1,172
en-kin	1,210
en-luo	1,199
en-twi	1,200
en-yor	1,206
fr-ewe	1,077
fr-wol	1,175
pt-nya	1,241
pt-vmw	930
Total	12,768

Table 1: Number of instances per language pair.

2020b), BLEU (Papineni et al., 2002), chrF (Popović, 2015), COMET22 (Rei et al., 2022a), COMETKiwi22 (Rei et al., 2022b), Sentinel-Cand (Perrella et al., 2024), Sentinel-Src (Proietti et al., 2025), spBLEU (Fan et al., 2020), and YiSi-1 (Lo, 2019).

3.2 Submissions

From each participating team, we received one primary submission and several secondary submissions. The primary submissions included MetricX-25 (Juraska et al., 2025), mr7.2.1 (Hrabal et al., 2025), Polycand-2 (Züfle et al., 2025), and rankedCOMET (Maharjan and Shrestha, 2025). The secondary submissions included baseCOMET, MetricX-25-QE, MetricX-25-Ref, Polycand-1, and Polyc-3.

MetricX-25 MetricX-25 is an encoder-only regression model initialized from Gemma3 12B (Team et al., 2025) and fine-tuned on WMT15–23 DA and MQM scores in two stages: first on z-normalized DA scores, then on a 1:1 mixture of rescaled DA and MQM scores with a score-type indicator to align outputs with the intended evaluation (ESA/DA vs. MQM). Compared to MetricX-24 (Juraska et al., 2024), this setup increases the weight of DA data in the second stage and explicitly conditions on score type. Both stages also include a small proportion of synthetic WMT-derived data used in MetricX-24. Training uses a maximum input length of 4K tokens to balance performance and coverage of low-resource language examples.

mr7.2.1 uses Gemma3 27B (Team et al., 2025), prompted with the DSPy framework and optimized with MIPROv2 (Opsahl-Ong et al., 2024), first generating seven aspect scores (0–10) and then producing the final overall score (0–100).

rankedCOMET is based on COMET22 (Rei et al., 2022a) used in zero-shot inference, producing raw segment-level scores that are then adjusted with per-language-pair rank normalization, yielding calibrated distributions and improved correlation with evaluation metrics.

Polycand-2 is a COMET-poly supervised model trained on WMT data up to 2024 (DA/ESA/MQM merged on a single scale). It extends COMET by using two alternative translations of the same source to provide better context for scoring. English–Korean and Japanese–Chinese were excluded from training.

3.3 AfriCOMET-1.1-MTL and SSA-COMET-MTL

For the WMT 2024 Metrics Shared Task on the African Challenge Set, the authors explored replacing the original AfroXLMR (Alabi et al., 2022) with an enhanced African-centric multilingual pretrained encoder, AfroXLMR-76L (Adelani et al.,

2024)[2], to build MT evaluation and QE models tailored for African languages, and named the resulting models as AfriCOMET-1.1 (Wang et al., 2024b). This upgrade yields notable improvements: AfriCOMET achieved the highest Spearman correlation on the 2024 WMT African Challenge Set among all benchmarked systems (unweighted setting), underscoring the critical role of a stronger base encoder in advancing the quality of MTE for African languages.

Since this challenge set was built on SSA-MTE, which provides a large-scale training set, we further explored the impact of incorporating in-domain, in-task training data on final model performance. Specifically, we report results for SSA-COMET-MTL (MTL stands for multi-task learning), currently the best SSA-COMET model, built using the same pipeline as AfriCOMET-1.1 with the same base model and training data, but augmented with additional training examples from SSA-MTE. As the original authors did not provide an MTL version of AfriCOMET-1.1, we reproduced it by following the same pipeline and hyperparameters, enabling a fair comparison. We chose to compare using the MTL versions of the models because, empirically, MTL models tend to outperform STL models when all other factors are held constant (Li et al., 2025).

4 Analysis

Table 2 shows the average segment-level correlations on the SSA-MTE challenge set. Overall, correlations are moderate, with most Spearman values between 0.3 and 0.5 and Pearson values between 0.35 and 0.55. We adapt the following definition of levels of agreement: a Spearman and a Pearson correlation lower than 0.4 indicates a low-level agreement, a value between 0.4 and 0.6 indicates medium-level agreement, and a value greater than 0.6 indicates high-level agreement with human judgments.

4.1 Official Baselines

Among the official baselines provided by the organizers, chrF and YiSi-1 achieve the strongest overall performance (Spearman 0.506 / 0.460, Pearson 0.532 / 0.493), indicating that carefully tuned lexical and embedding similarity measures remain competitive in this evaluation setting. Notably, chrF even has comparable performance to the best supervised participant submission, MetricX-25. spBLEU shows a notable advantage over BLEU in Spearman (0.434 vs. 0.389), suggesting better ranking stability when using sentencepiece tokenization for morphologically rich or orthographically diverse languages. COMET22 achieves a lower correlation (0.363 in Spearman) compared to AfriCOMET-1.1-MTL, which shares the same architecture but uses an African-language–enhanced encoder LM (AfroXLMR-76L). This mirrors last year's findings on the importance of the base model. COMETKiwi22, a reference-free system, exhibits a clear performance drop compared to its reference-present variant COMET22, indicating that reference information remains important for neural, learned metrics. The Sentinel metrics score lowest overall, suggesting that their coarse-grained features are insufficient for fine-grained adequacy judgments in this domain.

Metrics	Pearson	Spearman
Baseline		
chrF	0.532	0.506
YiSi-1	0.493	0.460
spBLEU	0.395	0.434
BERTScore	0.471	0.425
BLEU	0.336	0.389
COMET22	0.405	0.363
COMETKiwi22	0.253	0.244
sentinel-cand	0.107	0.102
sentinel-src	0.068	0.073
Primary		
MetricX-25	0.530	0.467
mr7.2.1	0.477	0.380
rankedCOMET	0.377	0.364
Polycand-2	0.142	0.132
Secondary		
MetricX-25-Ref	0.550	0.490
MetricX-25-QE	0.490	0.427
baseCOMET	0.405	0.364
Polyic-3	0.177	0.144
Polycand-1	0.159	0.152
Additional		
SSA-COMET-MTL	0.688	0.630
AfriCOMET-1.1-MTL	0.599	0.552

Table 2: Average segment-level correlation coefficients of MT evaluation metrics across languages on the SSA-MTE test set.

4.2 Participant Submissions

For participant submissions, the highest average correlations come from MetricX-25-Ref (Pearson 0.550, Spearman 0.490), outperforming both its QE variant (MetricX-25-QE) and its default vari-

[2] https://huggingface.co/Davlan/afro-xlmr-large-76L

ant (MetricX-25). This reinforces the finding that reference-based approaches are more effective than QE-only approaches in the SSA-MTE setting. MetricX-25 is the best among primary submissions, followed closely by MetricX-25-QE and rankedCOMET. mr7.2.1 is competitive in Pearson but drops in Spearman, while the Polycand/Polyic series performs notably lower, suggesting limited adaptation to the SSA-MTE adequacy signal.

4.3 Additional Baselines

We also include two other baselines intended to benchmark progress in African-language MT evaluation: AfriCOMET-1.1-MTL and SSA-COMET-MTL. Both use the same base model and pipeline, with SSA-COMET-MTL further incorporating in-domain SSA-MTE training data. SSA-COMET-MTL achieves the highest overall scores in this evaluation (Pearson 0.688, Spearman 0.630), outperforming AfriCOMET-1.1-MTL by +0.089 Pearson and +0.078 Spearman. These gains highlight the value of in-domain, in-task supervision. Among submitted systems, MetricX-25 and its variants achieve the highest correlations, generally reaching a medium-level agreement with human judgments. All other submissions remain in the low-agreement range (Spearman < 0.4), indicating notable room for improvement.

4.4 Per-LP Performance

Figure 1 presents per-language averages across all metrics. Performance varies substantially by language pair: pt-nya is the easiest (Pearson 0.606, Spearman 0.491), while pt-vmw is the hardest (Pearson 0.168, Spearman 0.147). English-sourced pairs show mixed difficulty, with en-amh and en-twi at the higher end, and en-luo and en-yor lower. Among French-sourced pairs, fr-wol tends to outperform fr-ewe. These trends suggest that both source–target pairing and specific linguistic features of the target language influence evaluation difficulty. However, only en-amh and pt-nya achieve medium-level Spearman correlations; all other LPs remain in the low-agreement range, with pt-vmw extremely low (around 0.1 on average).

In summary, correlations on the SSA-MTE challenge set remain moderate even for the strongest systems, underscoring the difficulty of MT evaluation for low-resource African languages. The results indicate that reference-based learned metrics with in-domain training (e.g., SSA-COMET-MTL, MetricX-25-Ref) offer clear advantages, but there is significant room for improvement before reaching high-agreement levels with human judgment. Promising directions include the use of African-language–enhanced encoders and leveraging large-scale in-domain, in-task training data.

5 Conclusion

We have presented the results of the WMT 2025 Metrics Shared Task for the SSA-MTE challenge set, the largest and most diverse benchmark to date for MT evaluation in Sub-Saharan African languages. The evaluation covered a wide range of metrics, including official baselines, participant submissions, and additional African-focused baselines. Overall, correlations with human judgments remain modest, with most submitted systems achieving Spearman correlations below 0.4, indicating low-level agreement. Reference-based neural metrics generally outperform reference-free approaches, with MetricX-25-Ref leading among participant systems. SSA-COMET-MTL, trained with in-domain SSA-MTE data, sets a new reference point for African-language MTE, demonstrating clear gains over AfriCOMET-1.1-MTL and underscoring the value of domain-matched supervision. Per-language analysis shows substantial variation in difficulty, reflecting differences in source–target pairing, linguistic complexity, and resource availability. Only en-amh and pt-nya surpass the 0.4 threshold for medium-level agreement, while pt-vmw remains particularly challenging with correlations near 0.1. These findings suggest that while progress has been made, significant room remains to improve robustness and accuracy for the most difficult language pairs. Future work should explore integrating African-language–enhanced encoders, expanding the diversity of training data, and developing methods that can better generalize across very low-resource languages.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius

Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Felermino D. M. A. Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. Building resources for emakhuwa: Machine translation and news classification benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14842–14857, Miami, Florida, USA. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. 22(1).

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. Cuni and phrase at wmt25 mt evaluation task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Ssa-comet: Do llms outperform learned metrics in evaluating mt for under-resourced african languages? *Preprint*, arXiv:2506.04557.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. Ranked-comet: Elevating a 2022 baseline to a top-5 finish in the wmt 2025 qe task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

NLLB-Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. *Preprint*, arXiv:2406.11695.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, and 39 others. 2024a. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024b. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516, Miami, Florida, USA. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. Comet-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

NVIDIA-NeMo's WMT 2025 Metrics Shared Task Submission

Brian Yan[1], **Shuoyang Ding**[2], **Kuang-Da Wang**[3],
Siqi Ouyang[1], **Oleksii Hrinchuk**[2], **Vitaly Lavrukhin**[2], **Boris Ginsburg**[2]
[1]Carnegie Mellon University, [2]NVIDIA, [3]National Yang Ming Chiao Tung University
byan@cs.cmu.edu, shuoyangd@nvidia.com

Abstract

This paper describes NVIDIA-NeMo's WMT 2025 Metrics Shared Task submission. We investigated two strategies for extending Machine Translation (MT) evaluation to unsegmented documents: 1) first segmenting into sentences and then applying regression-based metrics on aligned sentence pairs, and 2) directly utilizing the long-context capabilities of LLMs. The base comparison of the segmentation-based and LLM-based metrics on the WMT 2023-24 evaluation sets indicated that the former performs more robustly across language pairs. Thus we sought to improve the LLM-based approach by incorporating *relative evaluation* - this setting jointly evaluates all candidate translations at once and relative to each other, rather than evaluating each separately. Our experiments using the open-source Qwen3 LLM show that relative evaluation improves score correlations with human judgment, but only if the task is structured as a 2-stage evaluate-then-refine problem.

1 Introduction

For most of its history, machine translation (MT) research has revolved around the sentence as the primary unit of translation and evaluation. Standard MT benchmarks and evaluation protocols typically present systems with short, isolated segments, and assess their quality against human-produced references (Kocmi et al., 2024). This sentence-level paradigm has been driven by practical constraints: statistical and neural MT systems were long limited by both computational costs and modeling power (Kim et al., 2019). And in turn, automatic metrics, whether lexical (Papineni et al., 2002) or regression-based (Rei et al., 2020), were also designed to operate on sentence-like segments.

However, translation in real-world applications rarely occurs in isolation. Professional translators work with continuous documents, where meaning and style are shaped by discourse, context, and document-level coherence (Lommel et al., 2014). As neural MT systems have become more capable, particularly with the advent of large language models (LLMs), research has increasingly shifted toward document-level translation and evaluation (Zhu et al., 2024; Pang et al., 2025). This shift reflects both a practical need for evaluating translations as they appear in natural, unsegmented contexts and a growing recognition that many important aspects of quality, such as pronoun resolution, lexical consistency, and narrative flow, can only be judged when viewing the text holistically. This recent emergence of long-context MT models has made it technically feasible to translate much larger spans of text at once, raising a new research question in the field of MT evaluation: *is it better to apply existing sentence-level metrics to segmented text or apply LLM-based metrics directly to document-length text?*

In the WMT 2025 Metrics Shared Task, we focus specifically on the problem of unsegmented document evaluation: assessing the quality of MT outputs when the entire document is presented as a single sequence. Our submission compares two divergent strategies: (1) applying traditional regression-based metrics after segmenting documents into sentences, and (2) leveraging the long-context capabilities of LLMs to perform holistic document evaluation directly. To support our investigation, we simulated document-level human evaluations by concatenating the sentence-level MQM annotations from the WMT 2023 and 2024 Metrics Shared Tasks. Our experiments showed that the scaled up regression-based metric correlated better with human judgement than the LLM-based metric – we thus designed the former as our primary submission and the latter as our secondary submission to the WMT 2025 Metrics Shared Task.

While our primary findings favored the segmentation-based regression metric, we also explored ways to strengthen the LLM-based approach.

Proceedings of the Tenth Conference on Machine Translation, pages 920–925
November 8-9, 2025

In particular, we experimented with *relative evaluation* — a setup in which all candidate translations for a given source are presented to the model simultaneously and ranked against each other, rather than scored in isolation. We found that relative evaluation yielded only marginal improvements in correlation with human judgments, and only when implemented as a two-stage "evaluate-then-refine" process. Although these gains were not large enough to change our overall ranking of the two strategies, this line of work remains relevant for understanding how LLM prompting strategies interact with long-context MT evaluation. We therefore complement our main results with an analysis of why relative evaluation showed limited benefits and where it may still hold promise.

2 Document-level Human Evaluation

Prior to 2025, the WMT shared tasks were limited to the segment-level translation paradigm - all MT systems and all human evaluation operated on segments. To study document-level evaluation, we construct a simulated document-level MQM set by concatenating the source and target segments of each document and summing their MQM error counts. Table 1 compares the source lengths and MQM scores before and after document concatenation.

While this offers the best available proxy for document-level score correlations on WMT data, it has two notable limitations:

1. MT systems trained and evaluated on segmented inputs may differ from true document-level MT systems, which can exhibit distinct error patterns such as over- or under-translation.

2. MQM raters were instructed to consider surrounding context, but their judgments remained segment-focused; discourse-level phenomena may be underreported.

Note that we do not simulate document-level evaluation for the WMT 2024 sets because a large portion of segments were not evaluated. We also do not simulate document-level ESA data, as its 0–100 scale is not naturally additive, unlike MQM error counts. We therefore focus on the three language pairs from WMT 2023: En-De, He-En, and Zh-En.

Table 1: Comparison of segment and document-level WMT human evaluation data: average source character length (Len) and average MQM score (Score).

	Segment		Document	
Set	Len	Score	Len	Score
WMT23-En-De	354	-7.1	1028	-16.9
WMT23-He-En	84	-2.3	1719	-17.3
WMT23-Zh-En	40	-4.3	449	-25.1
WMT24-En-De	185	-3.0	-	-
WMT24-En-Es	185	-1.0	-	-
WMT24-Ja-Zh	91	-3.2	-	-

3 Metric Descriptions

3.1 Segmentation-based

The segmentation-based system breaks down document-level evaluation into sentence-level subproblems and relies on legacy sentence-level metrics for evaluation. Our approach first segments the source and target documents into sentences using ersatz (Wicks and Post, 2021), then establish aligned sentence blocks from these sentences using Vecalign (Thompson and Koehn, 2019) and LASER (Artetxe and Schwenk, 2019) sentence embeddings. Because our preliminary finding shows that existing sentence-level metrics are not good at identifying over- and under-translation errors, we have to rely on null alignments to identify these errors. To avoid merging unrelated sentences into aligned sentence blocks, we introduce an adaptive heuristic search strategy that dynamically finds the optimal alignment penalty (β_{skip}) for each document, ensuring those over- and under-translation errors are correctly identified as null alignments. Different from past practices such as mWERSegmenter (Matusov et al., 2005), which jointly align and segment system translations according to a reference, our system is a reference-free submission that directly aligns system translation to the source.

We apply MetricX-24-XXL-Hybrid-QE (Juraska et al., 2024) to each aligned blocks. In cases where null alignments are established, we penalize null alignments by assigning a score of 25 for each null alignment. We average the scores evalauted over sentence blocks to obtain a document-level score. To ensure the score aligns with the scale and directionality of WMT 2025, we apply a simple linear transformation of $100 - 4 \times s_d$ on the document-level score s_d.

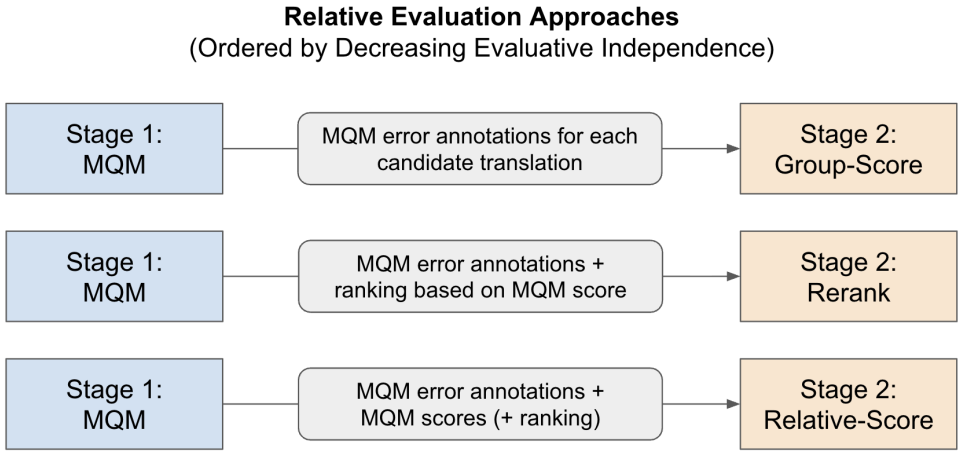


Figure 1: Overview of our 2-stage LLM-based methods: three relative evaluation approaches incorporating varying degrees of information from the initial MQM stage.

3.2 LLM-based

We implemented four document-level MT evaluation methods based on the Qwen3 large language model (LLM). All methods operate on unsegmented documents and follow the MQM (Multidimensional Quality Metrics) annotation scheme. The first method is a direct re-implementation of an established approach, while the remaining three extend it by incorporating a second-stage relative evaluation procedure.

3.2.1 Qwen-MQM

The baseline method, Qwen-MQM, is a Qwen-based re-implementation of the GEMBA-MQM framework, which was originally proposed using GPT models (Zhao et al., 2024). In this setting, the LLM is provided with the source document and a candidate translation and is prompted to produce MQM annotations (Kocmi and Federmann, 2023). The final document score is obtained by aggregating these annotations according to the MQM weighting scheme (Freitag et al., 2021). This method evaluates each system output independently, without reference to other candidate translations.

3.2.2 Relative Evaluation Extensions

The remaining three methods augment Qwen-MQM with a second evaluation stage in which the LLM is shown multiple system outputs for the same source document and instructed to re-assess or adjust scores based on cross-system comparison. This relative evaluation setting leverages the LLM's long-context capability to consider multiple translations simultaneously, with the goal of improving the utility of the metric for ranking MT systems. We present the methods in order of decreasing evaluative independence - that is, the degree to which final scoring decisions are formed without being conditioned on the scores and rankings of the initial MQM stage.

Qwen-MQM-Group-Score: In the second stage, the LLM is presented with the complete set of candidate translations together with their initial MQM annotations. It is then instructed to assign final scores to all candidates in a single pass, explicitly weighing the relative strengths and weaknesses that emerge through comparison. Since only the initial MQM annotations, and not the resultant MQM scores, are presented to the LLM in this second stage we consider that this approach has a high degree of evaluative independence.

Qwen-MQM-Rerank: In the second stage, the LLM is shown the initial ranking of candidates derived from their MQM scores, along with their MQM annotations, and is asked whether any adjustments to the ordering are warranted. Compared to the previous approach, this one offers less evaluative independence in the second stage: here, the initial ranking is presented to the LLM and the task is framed as a re-ranking of the initial judgments.

Qwen-MQM-Relative-Score: In the second stage, the LLM receives the initial MQM annotations and MQM scores (and thus implicitly the initial rankings) of all other candidates and is tasked with assigning a score to a single target translation in light of these scores. This creates the strongest dependency on prior judgments, as the evaluation of each candidate is explicitly anchored to the assessed quality of its competitors. Unlike the previ-

Table 2: Comparison of regression-based and LLM-based metrics on segment-level WMT data, as measured by system and segment score correlations with human judgments (Kendall's Tau). Correlations are averaged across language pairs.

Model	Sys	Seg
XComet-QE	0.731	0.371
MetricX-QE	0.769	**0.387**
GEMBA-MQM/ESA	**0.809**	0.362
Llama-MQM	0.674	0.230
Qwen-MQM	0.736	0.356

ous two approaches which evaluate all candidates at once, this one cycles through each candidate in order to evaluate each relative to the rest.

Figure 1 summarizes these three relative evaluation approaches in terms of what information from the initial MQM stage is made available. In other words, with more information (in the form of annotations, ranking, and scores), the evaluative independence of the second stage decreases.

4 Experimental Setup

Data: We use evaluation data from the WMT 2023–2024 Metrics Shared Tasks: MQM (both years) and ESA (2024). These are inherently segment-level corpora: translations were produced from pre-segmented sources, and human ratings were applied to individual segments.

Models: For LLM-based metrics, we use Qwen3-14B (Team, 2025b) in our experiments. The Qwen3 series is a hybrid reasoning/instruction-following model, capable of scaling inference-time compute via the generation of reasoning traces; however, we found that the reasoning mode degraded performance, so we disabled it in our experiments. In our attempt to identify the best open-source LLM, we also experimented with Llama3.1-8B (Grattafiori et al., 2024) and Gemma-3-12b-it (Team, 2025a) – these were less performant than Qwen3, as described in the next section.

Meta-Evaluation: For measuring score correlation with human judgment we rely on Kendall's Tau, as computed by the official WMT repository: `https://github.com/google-research/mt-metrics-eval`. We track both system-level and segment/document-level score correlations.

5 Results

5.1 Segment-level Evaluation

Table 2 compares regression-based and LLM-based metrics on the WMT 2023–2024 segment-level datasets. Correlations with human judgments are reported at both the system and segment level, averaged over all language pairs.

MetricX-QE stands out as the best overall metric, considering both system and segment-level score correlation. The GPT-based GEMBA metric is a close second, but for our submission we opted for open-source alternatives. Therefore on the LLM side, Qwen-MQM is the strongest available metric. We found that Llama produced reasonable, but weaker, results while Gemma failed to consistently follow the MQM instructions.

5.2 Document-level Evaluation

Given that MetricX-QE and Qwen-MQM were the strongest regression-based and LLM-based metrics respectively at the segment level, we centered our document level investigations around these two.

Table 3 presents results on the simulated document-level dataset constructed from WMT 2023 MQM annotations. Here the LLM-base metric outperformed the regression-based, unlike in the previous segment-level setting. This suggests that the long context capability of LLMs lead to a more holistic evaluation of document-level translations, while regression-based methods still require some form of segmentation into parts.

The single stage Qwen-MQM outperformed the two relative evaluation approaches with the highest degrees of evaluative independence in the second stage: Qwen-MQM-Rerank and Qwen-MQM-Group-Score. These degradations resulting from the second stage suggest that the relative evaluation ability of LLMs is still weak.

On the other hand, the Qwen-MQM-Relative-Score approach yielded moderate improvements over Qwen-MQM - since this approach provides a great deal of information (MQM annotations and MQM scores) from the first stage, it limits how much the scores produced in the second stage can deviate from that of the first stage.

6 Conclusion

We investigated two strategies for unsegmented document-level MT evaluation: scaling traditional regression-based metrics to longer contexts and applying LLM-based metrics capable of holistic

Table 3: Comparison of regression-based and LLM-based metrics on document-level WMT data, as measured by system and document score correlations with human judgments (Kendall's Tau).

Metric	En-De		He-En		Zh-En		Avg	
	Sys	Doc	Sys	Doc	Sys	Doc	Sys	Doc
LASER-MetricX-QE	0.909	0.328	0.848	0.233	0.714	0.285	0.824	0.282
Qwen-MQM	0.909	**0.429**	0.758	0.346	0.810	0.460	0.836	**0.425**
Qwen-MQM-Rerank	0.909	0.415	0.758	0.343	0.829	0.460	0.832	0.406
Qwen-MQM-Group-Score	**0.939**	0.365	**0.909**	0.250	**0.924**	0.261	**0.924**	0.292
Qwen-MQM-Relative-Score	**0.939**	0.421	**0.909**	**0.357**	0.810	**0.471**	0.886	0.416

assessment. While regression-based metrics exhibited stronger correlations with human judgments than LLM-based metrics on segment-level data, the inverse was true on simulated document-level data. While relative evaluation techniques modestly improved LLM-based performance, gains were only achieved under fairly restrictive settings. These findings suggest that long-context LLMs are a promising basis for document-level MT evaluation, but further work is needed to fully realize the potential of LLM-based approaches.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 24–34.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, and 1 others. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Evgeny Matusov, Gregor Leusch, Oliver Bender, and Hermann Ney. 2005. Evaluating machine translation output with automatic sentence segmentation. In *2005 International Workshop on Spoken Language Translation, IWSLT 2005, Pittsburgh, PA, USA, October 24-25, 2005*, pages 138–144. ISCA.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Gemma Team. 2025a. Gemma 3.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007, Online. Association for Computational Linguistics.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. Lora land: 310 fine-tuned llms that rival gpt-4, A technical report. *CoRR*, abs/2405.00732.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

GEMBA-MQM V2: Ten Judgments Are Better Than One

Marcin Junczys-Dowmunt
Microsoft

Abstract

We introduce GEMBA-MQM V2, an MQM-inspired, reference-free LLM evaluation metric for the WMT25 Metrics Shared Task (Subtask 1). Building on GEMBA/GEMBA-MQM, we prompt GPT-4.1-mini to produce structured MQM error annotations per segment. We map annotations to scores with 25/5/1 severity weights (minor punctuation = 0.1). To reduce stochastic variance, each segment is scored ten times and aggregated with a reciprocal-rank weighted average (RRWA) after removing outliers beyond 2σ. On the WMT24 MQM test sets, GEMBA-MQM V2 ranks first by average correlation, with strong results across languages and evaluation levels; WMT23 results show comparable performance.

1 Introduction

Automatic evaluation is essential for assessing machine translation quality at scale. Recent work shows that large language models (LLMs) can act as effective evaluators when guided by MQM-style prompts (Lommel et al., 2014; Kocmi and Federmann, 2023b,a). We revisit this approach for WMT25 and propose GEMBA-MQM V2 — a more robust extension of the original method.

Using GPT-4.1-mini and full source-document context, we obtain strong segment-level and competitive system-level correlations on WMT24 and WMT23, while controlling judgment variability via multi-run aggregation. Structured JSON inputs/outputs enable reliable parsing and a clean separation between prompt and payload.

2 Data and Evaluation Protocol

We use MQM human-annotated test sets from WMT23 and WMT24 (overviews: (Haddow et al., 2023, 2024); metrics tasks: (Freitag et al., 2023, 2024)) as distributed in mt-metrics-eval (Google Research, 2024), following the MQM standard (Lommel et al., 2014). For WMT24 we evaluate English–German (en–de), English–Spanish (en–es), and Japanese–Chinese (ja–zh); for WMT23 we use English–German (en–de), Hebrew–English (he–en), and Chinese–English (zh–en). We follow the official task scripts to compute system- and segment-level correlations and report the prescribed measures for each year and language pair.

As in the original GEMBA approach, we use GPT-4.1-mini without further training.

3 Prompts

GEMBA-MQM V2 prompts GPT-4.1-mini with: (a) an MQM system instruction defining severity (critical/major/minor) and error types, while providing full source-document context and (b) line-by-line JSON inputs carrying source, target, and language tags. The model returns a JSON object with lists of errors by severity and type, which we score as MQM "badness" with weights 25/5/1 (minor punctuation = 0.1), then negate so higher is better for mt-metrics-eval.

The MQM protocol for human annotators is based on error span annotation. Span marking is difficult for generative LLMs, so following Kocmi and Federmann (2023a) we elicit short error descriptions rather than spans. Scoring depends only on error severity and error type.

Figure 1 shows the system prompt and the full English context excerpt used in the example. We then process each text segment individually (split on newlines), as shown in Figure 2, which presents a concrete input/output JSON pair from a WMT24 document and one of its system outputs. Currently, we do not maintain the history of previous segments, only the system prompt is visible for each segment prompt. Relative to prior GEMBA-MQM, we switch to GPT-4.1-mini, enforce JSON outputs, set temperature to 0.4, and judge line-by-line with full document context present in the prompt.

Proceedings of the Tenth Conference on Machine Translation, pages 926–933
November 8-9, 2025

```
You are an annotator for the quality of machine translation. Your task is to identify errors and assess
the quality of the translation.

To accomplish this, you will receive a pair of paragraphs from this context as a JSON structure. For each
input, reply with an extended JSON object that contains the following information:

Focus on errors in the translation, not in the source. Each error is classified as one of three
categories: "critical", "major", and "minor". Critical errors inhibit comprehension of the text. Major
errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are
technically errors, but do not disrupt the flow or hinder comprehension.

For every of the main three categories, additionally identify error types in the translation and
sub-classify them. The types of errors are: "accuracy" ("addition", "mistranslation", "omission",
"untranslated text"), "fluency" ("character encoding", "grammar", "inconsistency", "punctuation",
"register", "spelling"), "style" ("awkward"), "terminology" ("inappropriate for context", "inconsistent
use"), "non-translation", or "other". For every error type, do also supply a short description ("desc") of
the error type. If there are no errors of a specific main category (critical, major or minor), it is OK to
return an empty list for that category. It also OK to not return any errors for any category if everything
is fine.

Here is an example of a JSON input (potentially other languages):
{
  "source_language": "English",
  "source": "I do apologise about this, we must gain permission from the account holder to discuss an
  order with another person, I apologise if this was done previously, however, I would not be able to
  discuss this with yourself without the account holders permission.",
  "target_language": "German",
  "target": "Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer
  anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die
  Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement."
}

And here is a corresponding JSON output with example error annotations (potentially other languages)
{
  "source_language": "English",
  "source": "I do apologise about this, we must gain permission from the account holder to discuss an
  order with another person, I apologise if this was done previously, however, I would not be able to
  discuss this with yourself without the account holders permission.",
  "target_language": "German",
  "target": "Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung mit einer
  anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber ohne die
  Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.",
  "errors": {
    "critical": [],
    "major": [
      {"type": "accuracy/mistranslation", "desc": "'involvement' is untranslated"}, {"type":
      "accuracy/omission", "desc": "'the account holder' is missing"}
    ],
    "minor": [
      {"type": "fluency/grammar", "desc": "'wäre' is a bit awkward"}, {"type": "fluency/register", "desc":
      "'dir' should be 'Sie'"}
    ]
  }
}

You should mimic the format from this example.

Apart from that, you are receiving the full English document as context which will help you analyze the
individual JSON segments provided after this for errors. Use this information to analyze the translation
pairs in their full context.

English input in context:
---
Siso's depictions of land, water center new gallery exhibition
"People Swimming in the Swimming Pool" from 2022 is one Vicente Siso artwork that will display at Tierra
del Sol Gallery beginning Jan. 13. (photo courtesy of Vicente Siso)
Tierra del Sol is pleased to present "Vicente Siso: Memories of the Land and Water" at the new gallery
location in West Hollywood. Siso has been an artist in the Studio Arts Program since 2012, this marks his
debut solo exhibition. Siso was born 1962 in Madrid and raised between Venezuela, Trinidad and Miami; he
moved with his family to Southern California in his early 20s.
Masterfully working across subject matter, Siso has generated a prolific series of landscapes, portraits,
and still-life works rendered in either acrylic, pastel, pencil or watercolor. Drawing from family
portraits, his own reference photographs, and recollection, his colorful compositions demonstrate his
range of interests and skill across media. Siso's tropical landscapes and seascapes reflect the
geographies of his past, employing rich patterns and incorporating people to make meaningful connections
between culture, memory and the environment. Siso titles his artworks in a mix of Spanish and English,
signifying the celebrated and integral complexities of his life in Los Angeles County. "Vicente Siso:
Memories of the Land and Water" opens on Saturday, Jan. 13, with a reception from 6-8 p.m. The exhibition
is on view through Sunday, March 3.
The Tierra del Sol Gallery is located at 7414 Santa Monica Blvd. For information, visit
tierradelsolgallery.org.
---

Next, you will receive the translation pairs as JSON objects. Analyze the translation as discussed above
and produce a JSON object with your analysis in response. Do not invent structural elements that are not
present in the JSON examples above. The only allowed keys are "source", "target", "source_language",
"target_language", "errors", "critical", "major", "minor", "type", "desc".
```

Figure 1: System prompt with JSON one-shot example and full document context

```
{
        "source_language": "English",
        "source": "Siso's depictions of land, water center new gallery exhibition",
        "target_language": "German",
        "target": "Sisos Darstellungen von Land, Wasserzentrum neuer Galerie"
}
```


(a) Input JSON object


```
{
        "source_language": "English",
        "source": "Siso's depictions of land, water center new gallery exhibition",
        "target_language": "German",
        "target": "Sisos Darstellungen von Land, Wasserzentrum neuer Galerie",
        "errors": {
                "critical": [
                        {
                                "type": "accuracy/mistranslation",
                                "desc": "The phrase 'water center new gallery exhibition' is
    mistranslated as 'Wasserzentrum neuer Galerie', which incorrectly suggests a 'water center'
    rather than 'water' being part of the exhibition's focus."
                        },
                        {
                                "type": "accuracy/omission",
                                "desc": "The word 'exhibition' is omitted; the translation ends
    with 'Galerie' (gallery) without indicating that it is an exhibition."
                        }
                ],
                "major": [],
                "minor": []
        }
}
```


(b) Output JSON object with annotation added (run 1 on WMT24 MSLC submission)

Figure 2: Follow-up prompt and output per text segment

4 Aggregation over multiple runs

Beyond prompt design, we consider aggregation across multiple runs to be the key contribution of GEMBA-MQM V2 for improving annotation reliability.

We score each segment ten times and aggregate the results using a reciprocal-rank weighted average (RRWA, see Section A.2): per segment, scores from each run are sorted in descending order, weighted by $1/\text{rank}$, and outliers beyond 2σ are removed before aggregation. This aggregation strategy is motivated by two observations:

- Individual runs can produce highly variable annotations due to the stochastic nature of the model.
- LLM-based MQM annotation often produces spurious error annotations.

We discuss these in more detail below.

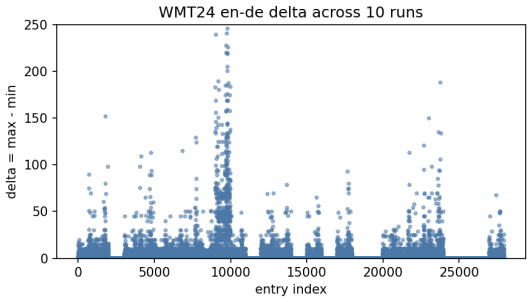
4.1 On variability

Individual runs with non-zero temperature produce highly variable annotations. After converting per-line outputs to MQM scores, we observe a substantial spread in values for the same segment. Table 1 illustrates this for the MSLC WMT24 example segment from the prompt: the same segment is judged ten times with very different outcomes.

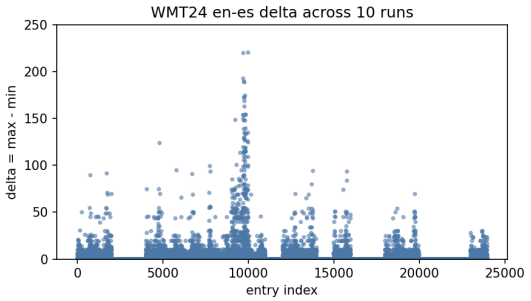
Run	Critical	Major	Minor	$> 2\sigma$	Score
1	2	0	0		-50
2	0	2	1		-11
3	0	1	1		-6
4	0	1	1		-6
5	0	2	1		-11
6	0	1	1		-6
7	0	1	1		-6
8	2	1	0	*	-55
9	0	1	1		-6
10	0	2	1		-11
mean-all		—			-16.85
mean		—			-12.55
max		—			-6.00
geo		—			-9.29
rrwa		—			**-8.50**

Table 1: Per-run MQM error counts and negated score for the segment from the MSLC prompt example. The outlier (*) is ignored. $\Delta = \max - \min = 49$.

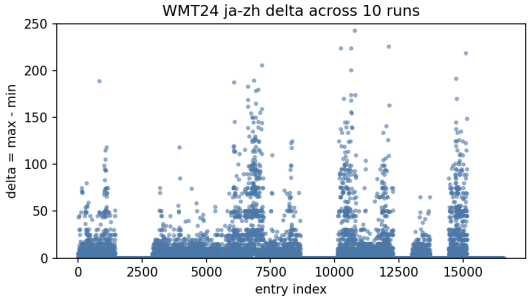
We further quantify variability by computing the difference (delta) between the maximum and minimum scores across ten runs for each segment. Figure 3 visualizes these deltas for all judged system outputs of WMT24 en–de, en–es, and ja–zh.



(a) en–de



(b) en–es



(c) ja–zh

Figure 3: Per-segment variability across 10 runs on WMT24. Each point shows $\Delta = \max - \min$; y-axis capped at 250. Zero-flat regions indicate systems without MQM gold, for which zeros were emitted.

Flat zero regions correspond to systems without MQM gold judgments; elsewhere, deltas can reach several hundred points, indicating substantial instability. Relying on a single judgment is risky.

4.2 On the tendency to over-annotate

LLM-based annotators often over-generate errors, sometimes identifying issues that are not present. Reciprocal-rank weighting biases the aggregate toward the lowest error magnitude (i.e., the most conservative plausible judgment), while remaining more discriminative than simply taking the minimum (or, for negated scores, the maximum). For comparison, we also report single-run variants (1–10), the simple mean, the geometric mean (geo), and the maximum (max). The reciprocal-rank weighted average (RRWA) consistently outperforms other aggregation methods, and all aggregates outperform individual runs (see Section 5 and Table 2).

Why not use the maximum then? One might expect the maximum score across runs (fewest errors after negation) to be the most conservative estimate. However, this can overlook subtle errors that are inconsistently annotated across stochastic runs. Since each annotation is a sample from a variable process, relying solely on the maximum risks ignoring genuine issues that appear in only some runs.

Aggregation, especially via reciprocal-rank weighting, balances caution with sensitivity to error diversity. MQM segment-level scores are weighted sums of few discrete values, often yielding repeated or tied scores (Table 1). Aggregating multiple runs increases score diversity and reduces ties, yielding a more nuanced and reliable estimate of translation quality. This better reflects the underlying variability in LLM-based annotation and mitigates the risk of over- or under-estimating errors. The aggregate thus behaves more like a regressed metric (e.g., MetricX-24).

4.3 On using other GPT variants and the lack of control

We observed challenges when using different GPT variants for evaluation. Although our approach relies on GPT-4.1-mini, other variants can produce markedly different behaviors and performance profiles. The lack of control over proprietary LLMs introduces variability across runs, model versions, and time periods.

During our experiments, we saw substantial behavior differences across time even when using what appeared to be the same model. Initially (December 2024), GPT-4o yielded results consistent with those reported here for GPT-4.1-mini. However, attempts to reproduce these results in May 2025 revealed significant performance drift for GPT-4o, rendering it unsuitable for our evaluation. GPT-4.1 showed similar degradation, whereas GPT-4.1-mini was performing well. We recovered our original December 2024 results only by reverting to an older, pinned version of GPT-4o.

Metric	Avg		en-de sys (pce)		en-de seg (acc-t)		en-es sys (pce)		en-es seg (acc-t)		ja-zh sys (pce)		ja-zh seg (acc-t)	
gemba-v2-gpt-4.1-mini-rrwa[noref]	1	0.728	16	0.829	1	0.550	4	0.823	2	0.688	1	0.921	3	0.557
MetricX-24	2	0.725	11	0.873	8	0.534	14	0.790	15	0.685	2	0.921	5	0.547
metametrics_mt_mqm_hybrid_kendall	3	0.724	5	0.882	6	0.542	9	0.803	9	0.686	15	0.871	2	0.561
metametrics_mt_mqm_kendall	4	0.724	6	0.881	4	0.542	7	0.803	7	0.686	14	0.871	1	0.561
metametrics_mt_mqm_same_source_targ	5	0.723	4	0.882	5	0.542	8	0.803	8	0.686	13	0.873	4	0.550
MetricX-24-Hybrid	6	0.721	10	0.873	9	0.532	11	0.798	13	0.685	5	0.896	7	0.539
XCOMET	7	0.719	1	0.906	10	0.530	15	0.789	3	0.688	7	0.889	12	0.510
MetricX-24-Hybrid-QE[noref]	8	0.714	9	0.879	12	0.526	13	0.792	16	0.685	10	0.875	8	0.530
gemba_esa[noref]	9	0.711	22	0.791	15	0.507	2	0.840	20	0.683	4	0.908	6	0.539
MetricX-24-QE[noref]	10	0.710	3	0.882	11	0.528	18	0.771	14	0.685	12	0.874	10	0.522
CometKiwi-XXL[noref]	11	0.703	15	0.839	21	0.481	1	0.843	35	0.680	9	0.881	14	0.494
XCOMET-QE[noref]	12	0.695	2	0.891	13	0.520	10	0.801	6	0.687	23	0.807	22	0.463
COMET-22	13	0.689	8	0.879	20	0.482	17	0.779	21	0.683	22	0.814	13	0.496
metametrics_mt_mqm_qe_same_source_t[noref]	14	0.688	12	0.858	17	0.497	20	0.710	11	0.686	18	0.852	9	0.524
BLEURT-20	15	0.686	7	0.881	19	0.486	22	0.696	26	0.681	8	0.887	18	0.484
metametrics_mt_mqm_qe_kendall.seg.s[noref]	16	0.684	13	0.858	18	0.497	21	0.710	12	0.686	20	0.838	11	0.516
bright-qe[noref]	17	0.681	21	0.817	16	0.500	12	0.794	1	0.689	24	0.805	17	0.484
BLCOM_1	18	0.665	14	0.843	23	0.455	24	0.682	25	0.681	19	0.842	16	0.488
sentinel-cand-mqm[noref]	19	0.650	19	0.821	14	0.517	16	0.787	19	0.683	31	0.610	19	0.481
PrismRefMedium	20	0.646	23	0.776	31	0.434	25	0.650	31	0.680	16	0.871	23	0.462
PrismRefSmall	21	0.642	24	0.772	33	0.433	26	0.632	34	0.680	11	0.875	25	0.457
CometKiwi[noref]	22	0.640	32	0.732	22	0.467	23	0.693	17	0.684	27	0.775	15	0.490
damonmonli	23	0.636	33	0.699	26	0.443	28	0.607	23	0.682	3	0.912	20	0.472
YiSi-1	24	0.630	25	0.762	29	0.436	27	0.609	28	0.681	21	0.835	24	0.458
monmonli	25	0.625	34	0.686	27	0.437	30	0.585	27	0.681	6	0.891	21	0.470
BERTScore	26	0.618	27	0.753	30	0.435	29	0.589	22	0.682	25	0.800	26	0.451
MEE4	27	0.609	31	0.733	28	0.437	35	0.500	18	0.683	17	0.857	27	0.446
chrF	28	0.608	26	0.753	35	0.431	31	0.582	37	0.680	28	0.766	32	0.436
chrfS	29	0.607	28	0.746	32	0.434	32	0.549	24	0.682	26	0.788	28	0.444
spBLEU	30	0.594	29	0.743	37	0.431	33	0.525	32	0.680	29	0.746	31	0.436
BLEU	31	0.589	30	0.737	36	0.431	34	0.514	36	0.680	30	0.736	36	0.435
BLCOM	32	0.537	35	0.615	34	0.433	19	0.731	33	0.680	34	0.327	33	0.435
XLsimDA[noref]	33	0.516	36	0.614	24	0.450	36	0.363	29	0.681	32	0.550	29	0.438
XLsimMqm[noref]	34	0.516	37	0.614	25	0.450	37	0.363	30	0.681	33	0.550	30	0.438
sentinel-ref-mqm	35	0.419	38	0.386	38	0.429	38	0.341	38	0.680	35	0.241	34	0.435
sentinel-src-mqm[noref]	36	0.419	39	0.386	39	0.429	39	0.341	39	0.680	36	0.241	35	0.435
gemba-v2-gpt-4.1-mini-2[noref]	8	0.723	25	0.819	7	0.541	5	0.838	11	0.687	16	0.904	9	0.551
gemba-v2-gpt-4.1-mini-3[noref]	9	0.723	16	0.836	14	0.535	14	0.804	16	0.686	2	0.922	6	0.555
gemba-v2-gpt-4.1-mini-8[noref]	11	0.723	29	0.813	10	0.539	2	0.840	8	0.687	12	0.910	12	0.548
gemba-v2-gpt-4.1-mini-7[noref]	12	0.722	23	0.821	11	0.538	6	0.829	20	0.685	8	0.915	17	0.544
gemba-v2-gpt-4.1-mini-4[noref]	13	0.722	26	0.817	16	0.534	11	0.819	23	0.685	1	0.924	8	0.551
gemba-v2-gpt-4.1-mini-9[noref]	14	0.721	30	0.808	9	0.539	8	0.827	21	0.685	4	0.921	15	0.546
gemba-v2-gpt-4.1-mini-1[noref]	16	0.720	17	0.831	15	0.535	13	0.808	5	0.688	10	0.913	14	0.547
gemba-v2-gpt-4.1-mini-10[noref]	17	0.719	24	0.820	8	0.539	12	0.810	15	0.686	13	0.909	7	0.553
gemba-v2-gpt-4.1-mini-5[noref]	19	0.716	28	0.816	13	0.537	19	0.801	7	0.687	15	0.904	11	0.549

Table 2: WMT24 with WMT24 task settings. Our GEMBA-MQM V2 variants compared to top systems.

5 Results on WMT24 metrics task data

We use the mt-metrics-eval toolkit to compute the correlations reported in Table 2. Our reference-free GEMBA-MQM V2 RRWA variant ranks first on WMT24 by average correlation (0.728), ahead of strong reference-based systems such as MetricX-24 (Juraska et al., 2024) and XCOMET (Guerreiro et al., 2024). Other reference-free metrics are further behind. As observed in the original GEMBA-MQM work (Kocmi and Federmann, 2023b,a), the strong performance of a general-purpose LLM is notable given that the competition includes purpose-trained metrics exposed to extensive task-specific human-created training data.

Single-run ablations group tightly (0.716–0.723), indicating good performance across stochastic runs despite high segment variability reported. Appendix Tables 3 and 4 provide WMT23 results under 2024 and original settings, respectively. Under 2024 rules our GEMBA-MQM V2 variant would have ranked first on WMT23 data as well. Aggregated results improve over each of the individual runs in every category. Segment-level performance is especially strong, while system-level performance lags behind. This suggests that the chosen MQM weights and the resulting segment scores may not lend themselves to cross-segment aggregation under equal weighting (simple mean).

6 WMT25 submission

For our WMT25 Subtask 1 submission, we follow the protocol outlined above. For all language pairs and systems we use the same prompts, temperature, and number of stochastic runs as in our WMT24 experiments.

The WMT25 Unified Evaluation task differentiates between language pairs with MQM-style scoring and Error Span Annotation (ESA) (Kocmi et al., 2024). We did not explore the implications of this differentiation in our current submission and simply submitted negated MQM scores for all language pairs. We expect rank-based correlations to carry over under this framework as in prior tasks.

At the time this paper was finalized, the shared task organizers had not yet released the final WMT25 metrics task results. As a consequence, we cannot report a definitive leaderboard position for our submission. Regretably, this reduces the value of this particular paper.

7 Conclusion

We presented GEMBA-MQM V2, a reference-free LLM evaluation metric/method for (machine) translation that combines JSON-first prompting, full source-document context, and multi-run aggregation. Scoring each segment ten times and aggregating with a reciprocal-rank weighted average (RRWA) improves robustness to stochastic variability and reduces over-annotation effects.

On WMT24 MQM test sets, GEMBA-MQM V2 ranks first by average correlation and is strong across languages and evaluation levels. Segment-level performance is especially strong, while system-level aggregation remains an open area for improvement.

References

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Chrysoula Zerva, and Alon Lavie. 2023. Results of the WMT23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 62–90, Singapore. Association for Computational Linguistics.

Google Research. 2024. MT metrics evaluation. GitHub repository. Accessed: 2025-08-13.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz. 2023. Findings of the WMT 2023 shared tasks. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–40, Singapore. Association for Computational Linguistics.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz. 2024. Findings of the WMT 2024 shared tasks. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–41, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Arle Richard Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and assessing translation quality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1160–1166, Reykjavik, Iceland. European Language Resources Association (ELRA).

Metric	Avg		en-de sys (pce)		en-de seg (acc-t)		he-en sys (pce)		he-en seg (acc-t)		zh-en sys (pce)		zh-en seg (acc-t)	
gemba-v2-gpt-4.1-mini-rrwa[noref]	1	0.760	2	0.977	7	0.597	8	0.948	7	0.566	4	0.920	3	0.549
GEMBA-MQM[noref]	2	0.754	1	0.982	18	0.572	10	0.947	11	0.563	1	0.937	27	0.522
MetricX-23-QE-b[noref]	3	0.753	6	0.961	1	0.606	15	0.939	3	0.576	26	0.896	7	0.539
XCOMET-Ensemble	4	0.751	17	0.939	3	0.604	18	0.936	1	0.584	22	0.901	5	0.543
MetricX-23-QE-c[noref]	5	0.750	15	0.946	12	0.581	21	0.930	6	0.572	2	0.927	4	0.545
XCOMET-XXL	6	0.749	18	0.938	4	0.603	14	0.940	4	0.575	23	0.900	6	0.541
MetricX-23-b	7	0.749	14	0.948	2	0.604	26	0.921	2	0.578	20	0.906	9	0.535
CometKiwi-XXL[noref]	8	0.742	3	0.970	14	0.578	33	0.911	24	0.550	11	0.917	22	0.528
XCOMET-XL	9	0.741	21	0.935	6	0.601	24	0.924	9	0.565	31	0.890	15	0.531
cometoid22-wmt23[noref]	10	0.741	16	0.945	10	0.586	19	0.933	28	0.540	6	0.920	28	0.520
MetricX-23	11	0.740	26	0.928	5	0.603	28	0.918	5	0.574	32	0.887	13	0.531
XCOMET-QE-Ensemble[noref]	12	0.737	22	0.934	9	0.588	25	0.922	18	0.552	30	0.892	11	0.533
CometKiwi-XL[noref]	13	0.735	4	0.968	19	0.571	36	0.905	30	0.533	16	0.914	26	0.522
MetricX-23-QE[noref]	14	0.734	25	0.929	8	0.596	30	0.914	12	0.561	34	0.876	23	0.527
COMET	15	0.729	7	0.960	16	0.574	20	0.931	33	0.530	36	0.868	32	0.514
MetricX-23-c	16	0.727	8	0.959	28	0.539	27	0.918	35	0.528	18	0.911	34	0.507
CometKiwi[noref]	17	0.727	19	0.938	20	0.569	41	0.889	27	0.543	25	0.896	25	0.525
mbr-metricx-qe[noref]	18	0.725	23	0.933	11	0.584	43	0.870	15	0.554	35	0.872	8	0.537
KG-BERTScore[noref]	19	0.722	20	0.936	24	0.556	40	0.892	29	0.536	27	0.894	30	0.516
BLEURT-20	20	0.721	9	0.956	17	0.572	38	0.902	36	0.517	38	0.863	29	0.518
docWMT22CometDA	21	0.718	5	0.964	23	0.559	17	0.936	44	0.491	37	0.863	41	0.493
docWMT22CometKiwiDA[noref]	22	0.717	10	0.955	25	0.547	34	0.909	45	0.484	12	0.917	40	0.493
cometoid22-wmt21[noref]	23	0.713	24	0.929	13	0.581	49	0.850	41	0.511	28	0.893	33	0.514
cometoid22-wmt22[noref]	24	0.713	28	0.925	15	0.578	48	0.852	39	0.513	29	0.893	31	0.515
instructscore	25	0.709	11	0.949	21	0.563	37	0.904	31	0.532	42	0.845	53	0.459
sescoreX	26	0.707	12	0.949	22	0.563	39	0.897	46	0.483	41	0.847	37	0.499
YiSi-1	27	0.706	30	0.915	27	0.542	32	0.911	32	0.530	45	0.835	35	0.504
MaTESe	28	0.705	38	0.870	31	0.528	31	0.912	26	0.546	24	0.898	47	0.479
Calibri-COMET22	29	0.701	31	0.906	35	0.522	12	0.945	40	0.513	43	0.844	49	0.474
prismRef	30	0.699	27	0.926	39	0.518	29	0.916	34	0.528	48	0.804	36	0.504
...														
gemba-v2-gpt-4.1-mini-6[noref]	6	0.752	2	0.979	14	0.584	3	0.953	23	0.550	15	0.914	16	0.530
gemba-v2-gpt-4.1-mini-9[noref]	7	0.752	8	0.976	26	0.577	1	0.960	20	0.551	10	0.917	21	0.528
gemba-v2-gpt-4.1-mini-4[noref]	9	0.751	12	0.970	22	0.578	2	0.956	25	0.550	3	0.921	19	0.529
gemba-v2-gpt-4.1-mini-2[noref]	10	0.751	10	0.975	21	0.579	5	0.951	22	0.550	13	0.916	10	0.534
gemba-v2-gpt-4.1-mini-1[noref]	11	0.751	11	0.974	13	0.585	6	0.949	16	0.554	17	0.912	18	0.529
gemba-v2-gpt-4.1-mini-5[noref]	14	0.749	16	0.967	27	0.577	4	0.952	13	0.555	9	0.918	20	0.528
gemba-v2-gpt-4.1-mini-10[noref]	15	0.749	5	0.977	20	0.580	9	0.948	14	0.555	21	0.903	14	0.531
gemba-v2-gpt-4.1-mini-3[noref]	17	0.748	7	0.976	19	0.580	22	0.929	17	0.554	5	0.920	12	0.531
gemba-v2-gpt-4.1-mini-7[noref]	18	0.747	14	0.970	18	0.581	13	0.940	19	0.552	14	0.914	24	0.526
gemba-v2-gpt-4.1-mini-8[noref]	19	0.747	4	0.977	25	0.577	16	0.939	21	0.550	19	0.907	17	0.529

Table 3: WMT23 evaluated with WMT24 task settings (retrofit protocol). We omitted systems with ranks above 30.

A Appendix

A.1 Results on WMT23

Tables 3 and 4 summarize WMT23 outcomes under the WMT24-retrofit and the original WMT23 protocols, respectively. Under the 2024 rules, our GEMBA-MQM V2 variant mirrors the WMT24 behavior and would have ranked first by average correlation. Under the original 2023 settings, results remain strong and consistent. In both protocols, multi-run aggregation improves over individual runs; segment-level performance is especially strong, while system-level performance lags behind, motivating future work on cross-segment aggregation beyond simple means.

A.2 Reciprocal-rank weighted average

Let $\{s_i\}_{i=1}^{k}$ be the per-run segment scores (higher is better; we use negated MQM). After removing outliers beyond 2σ (Section 4), sort the remaining n scores in descending order $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(n)}$. The reciprocal-rank weighted average (RRWA) is

$$\mathrm{RRWA}(\{s_i\}) = \frac{\sum_{r=1}^{n} w_r \, s_{(r)}}{\sum_{r=1}^{n} w_r}, \quad \text{with} \quad w_r = \frac{1}{r}.$$

Thus, higher-ranked (larger) scores receive larger weights, biasing the aggregate toward more conservative judgments while remaining sensitive to mid/low ranks.

Metric	all Avg		all sys acc		en-de sys		en-de seg		he-en sys		he-en seg		zh-en sys		zh-en seg	
XCOMET-Ensemble	1	0.825	7	0.928	10	0.980	2	0.695	17	0.950	1	0.556	27	0.927	2	0.650
XCOMET-XXL	2	0.824	5	0.932	8	0.982	1	0.695	12	0.964	2	0.556	29	0.911	5	0.627
MetricX-23-QE-b[noref]	3	0.823	2	0.940	9	0.982	5	0.628	18	0.947	6	0.529	28	0.926	1	0.661
XCOMET-XL	4	0.816	8	0.924	19	0.973	3	0.680	24	0.937	5	0.536	39	0.884	7	0.624
gemba-v2-gpt-4.1-mini-rrwa[noref]	5	0.814	6	0.928	6	0.988	7	0.597	8	0.967	12	0.463	11	0.956	12	0.562
MetricX-23-QE-c[noref]	6	0.813	4	0.932	21	0.972	12	0.525	21	0.939	7	0.526	4	0.974	10	0.581
MetricX-23-b	7	0.811	10	0.916	4	0.990	10	0.566	28	0.928	4	0.537	32	0.902	8	0.612
XCOMET-QE-Ensemble[noref]	8	0.808	14	0.908	17	0.974	4	0.679	36	0.909	9	0.498	36	0.892	4	0.647
MetricX-23	9	0.808	13	0.908	13	0.977	8	0.585	35	0.910	3	0.548	41	0.873	6	0.625
GEMBA-MQM[noref]	10	0.802	1	0.944	1	0.993	17	0.502	22	0.939	26	0.401	1	0.991	29	0.449
MetricX-23-QE[noref]	11	0.800	25	0.892	23	0.969	6	0.626	48	0.858	8	0.520	43	0.859	3	0.647
cometoid22-wmt23[noref]	12	0.794	3	0.936	11	0.979	21	0.448	29	0.928	29	0.397	10	0.956	32	0.439
mbr-metricx-qe[noref]	13	0.788	30	0.880	14	0.976	9	0.571	32	0.915	25	0.411	24	0.936	26	0.489
CometKiwi-XXL[noref]	14	0.786	12	0.912	7	0.986	29	0.417	27	0.929	32	0.390	2	0.978	37	0.390
CometKiwi-XL[noref]	15	0.786	9	0.916	15	0.975	22	0.446	42	0.900	35	0.384	3	0.974	34	0.430
MaTESe	16	0.782	18	0.904	37	0.918	11	0.554	38	0.906	15	0.459	38	0.889	24	0.511
CometKiwi[noref]	17	0.782	17	0.904	28	0.946	19	0.475	47	0.860	33	0.387	6	0.963	31	0.442
COMET	18	0.779	21	0.900	3	0.990	26	0.432	20	0.940	27	0.401	34	0.898	35	0.396
MetricX-23-c	19	0.778	11	0.916	29	0.944	16	0.508	19	0.946	44	0.313	17	0.953	40	0.371
instructscore	20	0.777	23	0.896	26	0.952	13	0.519	34	0.910	16	0.458	44	0.825	25	0.499
BLEURT-20	21	0.776	24	0.892	5	0.990	18	0.484	25	0.937	37	0.382	40	0.880	39	0.378
KG-BERTScore[noref]	22	0.774	28	0.884	31	0.926	20	0.451	37	0.908	36	0.382	7	0.962	33	0.430
sescoreX	23	0.772	26	0.892	27	0.952	14	0.519	41	0.901	34	0.385	48	0.797	17	0.536
cometoid22-wmt22[noref]	24	0.772	29	0.880	18	0.973	24	0.441	50	0.839	39	0.365	23	0.940	27	0.479
cometoid22-wmt21[noref]	25	0.768	31	0.871	20	0.973	27	0.428	51	0.832	40	0.360	26	0.929	28	0.458
docWMT22CometDA	26	0.768	19	0.904	2	0.990	31	0.394	30	0.922	41	0.339	30	0.907	42	0.353
docWMT22CometKiwiDA[noref]	27	0.767	22	0.900	22	0.970	23	0.444	39	0.906	46	0.286	5	0.965	38	0.387
Calibri-COMET22	28	0.767	16	0.904	24	0.963	30	0.413	26	0.930	28	0.401	42	0.863	36	0.396
Calibri-COMET22-QE[noref]	29	0.755	35	0.863	12	0.978	25	0.441	53	0.778	31	0.395	25	0.934	30	0.443
YiSi-1	30	0.754	34	0.871	32	0.925	32	0.366	31	0.917	30	0.395	45	0.823	43	0.290
...																
gemba-v2-gpt-4.1-mini-10[noref]	11	0.808	6	0.932	20	0.982	11	0.576	6	0.970	14	0.462	15	0.954	22	0.529
gemba-v2-gpt-4.1-mini-6[noref]	13	0.808	9	0.932	2	0.991	12	0.574	1	0.973	24	0.437	9	0.957	20	0.531
gemba-v2-gpt-4.1-mini-4[noref]	14	0.807	7	0.932	14	0.986	19	0.560	5	0.970	13	0.462	8	0.958	18	0.536
gemba-v2-gpt-4.1-mini-5[noref]	15	0.807	8	0.932	9	0.988	17	0.560	4	0.970	18	0.457	20	0.950	21	0.529
gemba-v2-gpt-4.1-mini-2[noref]	16	0.806	10	0.928	18	0.983	15	0.566	3	0.971	21	0.446	21	0.950	16	0.539
gemba-v2-gpt-4.1-mini-9[noref]	17	0.806	11	0.928	17	0.984	18	0.560	10	0.967	19	0.453	19	0.952	13	0.545
gemba-v2-gpt-4.1-mini-1[noref]	18	0.805	17	0.924	12	0.987	23	0.550	11	0.965	23	0.439	18	0.953	14	0.544
gemba-v2-gpt-4.1-mini-3[noref]	19	0.805	18	0.924	8	0.988	21	0.553	13	0.959	20	0.447	12	0.955	15	0.540
gemba-v2-gpt-4.1-mini-7[noref]	20	0.805	19	0.924	13	0.986	22	0.553	2	0.973	17	0.457	16	0.954	19	0.532
gemba-v2-gpt-4.1-mini-8[noref]	21	0.803	20	0.924	16	0.985	16	0.562	14	0.959	22	0.444	22	0.948	23	0.521

Table 4: WMT23 with WMT23 task settings (as originally reported). We omitted systems with ranks above 30.

Example using Table 1: after removing the single outlier (-55), the sorted scores are

$$\{-6,\ -6,\ -6,\ -6,\ -6,\ -11,\ -11,\ -11,\ -50\}.$$

With $w_r = 1/r$ and $\sum_{r=1}^{9} w_r \approx 2.83$, the numerator is

$$-\tfrac{6}{1} - \tfrac{6}{2} - \tfrac{6}{3} - \tfrac{6}{4} - \tfrac{6}{5} - \tfrac{11}{6} - \tfrac{11}{7} - \tfrac{11}{8} - \tfrac{50}{9} \approx -24.04.$$

Hence,

$$\mathrm{RRWA} \approx \frac{-24.04}{2.83} \approx -8.50,$$

matching Table 1. In our setup, RRWA acts like a soft maximum: it heavily favors the best (least-error) runs while still allowing the remainder to pull the score down when multiple runs consistently find issues. With 10 runs, the cumulative mass on the top ranks is substantial: top-1 $\approx 34\%$, top-2 $\approx 51\%$, top-3 $\approx 63\%$, and top-5 $\approx 78\%$.

CUNI and Phrase at WMT25 MT Evaluation Task

Miroslav Hrabal,[1] Ondřej Glembek,[2] Aleš Tamchyna,[2] A. Silja Hildebrand,[2] Alan Eckhardt,[2] Miroslav Štola,[2] Sergio Penkale,[2] Zuzana Šimečková,[2] Ondřej Bojar,[1] Alon Lavie,[2] Craig Stewart[2]

[1]Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{hrabal,bojar}@ufal.mff.cuni.cz

[2]Phrase a.s.
name.surname@phrase.com

Abstract

This paper describes the joint effort of Phrase a.s. and Charles University's Institute of Formal and Applied Linguistics (CUNI/UFAL) on the WMT25 Automated Translation Quality Evaluation Systems Shared Task. Both teams participated both in a collaborative and competitive manner, i.e. they each submitted a system of their own as well as a contrastive joint system ensemble. In Task 1, we show that such an ensembling—if chosen in a clever way—can lead to a performance boost. We present the analysis of various kinds of systems comprising both "traditional" NN-based approach, as well as different flavours of LLMs—off-the-shelf commercial models, their fine-tuned versions, but also in-house, custom-trained alternative models. In Tasks 2 and 3 we show Phrase's approach to tackling the tasks via various GPT models: Error Span Annotation via the complete MQM solution using non-reasoning models (including fine-tuned versions) in Task 2, and using reasoning models in Task 3.

1 Introduction

Machine translation (MT) evaluation has evolved rapidly in recent years, driven by the dual rise of large language models (LLMs) and specialised neural quality estimation (QE) architectures. Shared tasks, such as the WMT MT Evaluation campaign, provide a rigorous and reproducible framework for assessing advances in automated evaluation. The WMT25 MT Evaluation Task continues this tradition, offering a benchmark for comparing diverse approaches across multiple subtasks, from system-level quality prediction to fine-grained error annotation.

In this paper, we present a joint study by Phrase a.s. and Charles University's Institute of Formal and Applied Linguistics (CUNI/UFAL), combining independent research streams with collaborative experimentation. Both teams participated in Task 1, each submitting their own primary and secondary systems, as well as a contrastive joint ensemble in Task 1. This setting allowed us to examine not only the relative strengths of distinct modelling paradigms but also the potential synergies of cross-team integration. Phrase also participated in Task 2 and Task 3.

Our contributions are threefold. First, we investigate complementary modelling strategies for Task 1, including NN-based regression models trained on large-scale MQM-style data (multidimensional quality metrics, Lommel et al., 2013), and LLM-based systems producing full linguistic quality annotations. We analyse how fusing/ensembling these approaches, when carefully configured, can yield measurable performance gains. Second, for Task 2, we evaluate GPT-based AutoLQA (Automatic Language Quality Assessment) systems—both off-the-shelf and fine-tuned variants—demonstrating their applicability to error span annotation in MQM and ESA (error span annotation, Kocmi et al., 2024) formats. Third, for Task 3, we explore reasoning-enabled LLMs with targeted prompting for translation improvement, highlighting trade-offs between fluency- and accuracy-oriented objectives.

By documenting both competitive and collaborative results, we aim to provide insight into practical design choices—such as metric selection, data sampling, and system fusion—that influence performance in automated MT evaluation. Beyond the leaderboard standings, our analysis identifies patterns that may inform future research on hybrid evaluation architectures, the role of fine-tuning in LLM-based evaluators, and the limits of current metrics in capturing subtle quality differences.

Proceedings of the Tenth Conference on Machine Translation, pages 934–944

2 Task 1

In Task 1, Phrase and UFAL decided to collaborate and choose the strategy in which each organisation submits its own primary system and one of its own secondary systems. Each team dedicated the other secondary slot to a joint system ensemble. The submitted systems are marked with an asterisk* in the respective tables.

2.1 Phrase Systems

We have two sets of models that we experimented with in Task 1: (1) NN-based quality regression models that directly output the quality prediction score, and (2) LLM-based models that run the full linguistic quality assessment that is used to calculate the final score, as described below.

There are two important decisions that we made in the course of the evaluations mainly due to our internal research conventions and momentum, and also due to some technical difficulties.

The first decision was about the development metric for Task 1. Although acc^*_{eq} *tie-calibrated pairwise accuracy* (Deutsch et al., 2023) is the primary metric of the WMT25 evaluations, we used Kendall's τ and Pearson r coefficients as our development proxy metrics. The main reason for this decision was our late adoption and implementation issues of the acc^*_{eq} metric. Note that we also used the two proxy metrics when fusing the CUNI and the Phrase systems.

The second decision was about the submission score. We decided to treat all scores in a unified MQM/ESA fashion, i.e. internally we do not distinguish between the two scores. The first reason for this adoption is that the evaluation is insensitive to the dynamic range of the two metrics and—to our knowledge—the difference between the two is (1) lower penalization of the "Fluency/Punctuation" MQM category and (2) higher penalization of the non-translations. The second reason is rather practical: at Phrase, our tools (that we used off-the-shelf for this evaluation) internally work with a single score (let us refer to the score merely as MQM) that is defined as follows: Given the input segment $\mathcal{X}$ (with its source and hypothesis fields), we compute the MQM score as:

$$\mathrm{MQM}(\mathcal{X}) = \max\left\{0, 1 - \frac{\mathrm{pen}\,(\mathcal{X})}{|\mathcal{X}_{\mathrm{source}}|}\right\}, \quad (1)$$

where $|\cdot|$ denotes the word count (note that we compute the word count of the source segment), and $\mathrm{pen}(\cdot)$ is the sum of all penalty points retrieved from the AutoLQA system output, where we map the minor, major and critical errors to the penalties of 1, 5 and 5, respectively (i.e., no quantitative difference between major and critical errors).

2.1.1 Phrase Development Data

As development data we used the Google-MQM (Freitag et al., 2021) dataset and the WMT22-QE combined dev and test sets. Since the latter only covers three language pairs (`en-de`, `en-ru` and `zh-en`, 1500 segments each), we also pulled test data from the Google-MQM set, which covers seven language pairs (`en-de`, `en-es`, `en-ru`, `en-zh`, `he-en`, `ja-zh`, `zh-en`). We randomly selected 700 segments per LP while making sure that only one system's translation of a source sentence was included. We then removed all translation variants of those source sentences from the remaining data to prevent contamination and filtered out very self-similar segments as well, resulting in a training dataset of about 150k segments.

2.1.2 Phrase QPS Systems

One set of features for the ensembles is provided by research variants of the Phrase QPS (Quality Performance Score) system[1][2] (*pqps*). We experimented with three pre-trained models: XLMR-large (Conneau et al., 2019), GTE (Zhang et al., 2024), and RemBERT (Chung et al., 2021), which we selected not only according to their accuracy, but also with regard to memory and inference speed requirements for the production environment at Phrase.

We continued training on a large amount of internal Phrase post-edit data covering 226 language pairs with chrF (Popović, 2015) as gold-label (resulting models are tagged *pe* in the result tables.)

Then we further fine-tuned on combined data from internal Phrase MQM-annotated and post-edit data as well as our training split of the Google-MQM and Amazon-bio-MQM datasets (Zouhar et al., 2024). We experimented with three gold/reference score variants: a balanced mix of MQM-score and chrF labels (*bal*), the same mix with a smoothed version of the MQM score (*balsm*) and one variant with only MQM- and MQM-like-scores (*mqm*) as gold label. In all cases, Equa-

[1] https://phrase.com/phrase-quality-technologies/quality-performance-score/

[2] https://support.phrase.com/hc/en-us/articles/5709672289180-Phrase-QPS-Overview

tion (1) was the fundamental instrument for the MQM gold label score calculation.

See Table 1 for individual system performance on our development test sets (see Section 2.1.1). Due to the limited number of slots, we did not submit the best Phrase QPS system.

2.1.3 Phrase AutoLQA System

The AutoLQA system (Automatic Linguistic Quality Assessment) is an LLM-based system whose underlying model is a fine-tuned `gpt-4o-mini` model, marked as `gpt-4o-mini-FT`. This is the very same system that was used for the primary submission of Task 2 (see Section 3 for reference). The system produces complete MQM annotations based on which the MQM-score is produced via Equation (1).

This system is referred to as *palqa-wmt25* in Table 1.

2.2 CUNI/UFAL Systems

This section describes the systems submitted by the CUNI team.

2.2.1 Training and Development data

For training, we used a subsample of WMT24 ESA scores, counting only 2160 examples in total.

Because high ESA scores are much more frequent in the full dataset, we resampled the data to produce a more uniform score distribution, while ensuring an even spread across language pairs.

As the development set, we used WMT23 metrics task en2cs DA scores. Due to the computational cost of the evaluation, we did not use other language pairs for the evaluation during development.

2.2.2 Systems based on Gemma 3 27B-it and DSPy

Our primary submission (*mr7.2.1*, in Table 2) and one of the secondary submissions (*mr6*, in Table 1 and Table 2) were based on the Gemma 3 27B-it[3] used via the DSPy framework (Khattab et al., 2024). Its MIPROv2 optimiser (Opsahl-Ong et al., 2024) was used to automatically select n-shot examples and adjust the prompts for improved performance.

[3]For both optimisation and inference, we run several instances of the model using vLLM 0.9.2+rocm641, each instance on 2x AMD MI210 with 16k tokens context. We used a LiteLLM proxy server for load balancing between instances. We also used GNU Parallel (Tange, 2025) as a workaround for performance issues when trying to scale the number of concurrent requests to fully utilise GPUs.

We selected the Gemma 3 27B-it model for its relatively compact size and strong multilingual capabilities. Our focus was on using reasonably sized open-weight models that we can run on our hardware, so we chose not to use larger commercial models available through APIs, even though we expect them to perform better.

The optimisation target was essentially rescaled Mean Squared Error:

$$\text{score}(X) = \frac{1}{n}\sum_{i=1}^{n} 1 - \frac{(X_{i,\text{gold}} - X_{i,\text{pred}})^2}{100^2} \quad (2)$$

Both submissions first predict,in one or multiple LLM calls, seven integer scores (0–10) covering different translation quality dimensions:

- accuracy and completeness
- terminology and consistency
- fluency and coherence
- style tone and audience fit
- locale conventions and formatting
- technical integrity
- cultural appropriateness

These categories were inspired by MQM, with initial detailed descriptions drafted by GPT o3-mini, see Appendix B for details on the optimized prompts:

After predicting these dimensions, both *mr6* and *mr7.2.1* output an overall translation quality score (0–100 integer). Similarly to the approach taken by the Phrase team, we chose not to distinguish between pairs that solicit ESA scores and MQM scores and use this 0-100 score for both.

The key difference between the systems lies in the overall score aggregation:

- **mr6:** Predicts each dimension separately in isolation, with a separate request for each. The overall score is then predicted in a separate LLM call, using only the dimension scores (not the original text). This approach was motivated by the lack of gold data for dimension-level scores and the absence of a clear formula to combine them. We originally planned to replace this aggregation with a simpler regression model (e.g., linear regression or gradient

Table 1: Task 1 Individual models' performance. The asterisk * denotes submitted systems: *palqa-wmt25* is Phrase's secondary submission and *cuni-mr6-overall* is CUNI's secondary submission. Best scores in each group in bold.

		google-mqm		wmt22-qe	
Feature/Model	Type	Kendall's τ	Pearson r	Kendall's τ	Pearson r
pqps-gte-pe	NN	0.0776	0.1093	0.2185	0.3710
pqps-rembert-pe	NN	0.0970	0.1486	0.2996	0.4576
pqps-xlmr-pe	NN	0.0954	0.1401	0.2760	0.4298
pqps-gte-bal-v1	NN	0.2306	0.4538	0.2670	0.4293
pqps-gte-bal-v2	NN	0.2324	0.4547	0.2589	0.4349
pqps-rembert-bal	NN	0.2465	0.4855	0.3184	0.4830
pqps-xlmr-bal	NN	0.2568	0.4936	0.3083	0.4778
pqps-gte-balsm	NN	0.2295	0.3973	0.2554	0.4390
pqps-rembert-balsm	NN	0.2422	0.4106	**0.3235**	**0.4945**
pqps-xlmr-balsm-v1	NN	0.2164	0.4288	0.2491	0.4416
pqps-xlmr-balsm-v2	NN	0.2292	0.3736	0.2993	0.4804
pqps-gte-mqm	NN	0.2398	0.4660	0.0789	0.2828
pqps-rembert-mqm	NN	0.2406	0.4652	0.2572	0.4119
pqps-xlmr-mqm	NN	**0.2620**	**0.5110**	0.0058	0.0954
palqa-wmt25*	LLM	**0.2790**	**0.4987**	**0.3145**	**0.4017**
cuni-mr6-y0	LLM	0.1767	0.3058	0.2153	0.2308
cuni-mr6-y1	LLM	0.1794	0.3311	0.2176	**0.2851**
cuni-mr6-y2	LLM	0.1800	0.2711	0.2240	0.2008
cuni-mr6-y3	LLM	0.1923	0.3098	0.2139	0.2192
cuni-mr6-y4	LLM	0.1249	0.2085	**0.2332**	0.2297
cuni-mr6-y5	LLM	0.1647	0.2759	0.1886	0.1978
cuni-mr6-y6	LLM	**0.2026**	0.3386	0.2034	0.2259
cuni-mr6-overall*	LLM	0.1826	**0.3460**	0.2214	0.2481

boosting) after the initial MIPROv2 optimisation, but our preliminary experiments showed no consistent improvements. We therefore submitted the LLM-based aggregation version, although the lack of gain might be due to unidentified implementation issues.

- **mr7.2.1:** Predicts all dimensions and the overall score in a single LLM call.

Both systems use the chain-of-thought technique (Wei et al., 2022) as implemented by the DSPy.[4] If the model fails to produce the response in the correct format, there are 2 retries with different temperatures: 0.5 and $\frac{2}{3}$.

In addition, *mr7.2.1* includes a fallback mode that bypasses chain-of-thought and generates the final answer directly if reasoning fails in all 3 tries (most commonly due to getting stuck in generating repetitive loops).

The complete training and inference code for the models submitted by CUNI will be available on GitHub.[5]

We evaluated our systems on WMT23 metrics using the mt-metrics-eval tool. We show the results in Table 2.

To get some idea whether the MIPROv2 optimisation and chain-of-thought actually contributes to better performance, we evaluate several modifications of the *mr7.2.1* submission: *_noopt* is a version with no MIPROv2 optimisation and consequently no n-shot examples. *_nocot* is a version where the output is generated directly with no chain-of-thought reasoning. The *_nocot_noopt* variant disables both. We can notice that disabling either of these or both at the same time seems to worsen the performance in all of the tracked metrics.

[4] https://dspy.ai/api/modules/ChainOfThought/

[5] https://github.com/hrabalm/wmt25-mt-eval-task

Table 2: CUNI models' performance on WMT23 en2cs dev set: segment-level Pearson r, system-level Pearson r and segment-level accuracy with optimised tie threshold. Our systems are in bold, systems marked with ** are CUNI primary submissions and systems marked with * are CUNI secondary submissions.

Model	Seg. P r	Rank	Sys. P r	Sys. rank	acc^*_{eq}	acc^*_{eq} rank
XCOMET-Ensemble	0.4017	1	0.9025	1	0.5404	2
XCOMET-QE-Ensemble[noref]	0.3952	2	0.9082	1	0.5286	4
COMET	0.3771	3	0.8646	2	0.5239	5
BLEURT-20	0.3730	3	0.7935	3	0.5111	7
MetricX-23	0.3606	4	0.8914	1	0.5500	1
KG-BERTScore[noref]	0.3503	4	0.7899	3	0.5062	10
CometKiwi[noref]	0.3503	5	0.7898	3	0.5171	6
MetricX-23-QE[noref]	0.3477	5	0.8782	2	0.5392	3
cometoid22-wmt22[noref]	0.3411	6	0.8254	3	0.5012	11
mr7.2.1[noref]**	0.3320	6	0.8021	3	0.3858	25
mr7.2.1_noopt[noref]	0.3146	7	0.7452	4	0.3351	27
mr7.2.1_nocot[noref]	0.3144	7	0.7508	4	0.3483	26
mr6[noref]*	0.3142	7	0.8114	3	0.4078	24
mr7.2.1_nocot_noopt[noref]	0.3115	7	0.7494	4	0.3295	29
GEMBA-MQM[noref]	0.3094	7	0.8520	2	0.3295	28
MS-COMET-QE-22[noref]	0.2864	8	0.7965	3	0.4984	12
prismRef	0.2649	9	0.5571	5	0.4910	14
XLsim	0.2589	9	0.6268	4	0.5075	9
YiSi-1	0.2453	10	0.5677	4	0.4968	13
BERTscore	0.2283	11	0.4798	5	0.4899	15
tokengram_F	0.2031	12	0.4087	7	0.4817	17
chrF	0.2006	13	0.4495	6	0.4794	18
f200spBLEU	0.1986	13	0.4962	5	0.4793	19
BLEU	0.1857	14	0.5186	5	0.4612	23
embed_llama	0.1720	15	0.4661	6	0.4767	20
prismSrc[noref]	0.1710	15	-0.0416	7	0.4615	22
eBLEU	0.1689	15	0.4672	6	0.4837	16
mre-score-labse-regular	0.1298	16	0.7184	4	0.5089	8
Random-sysname[noref]	0.0018	17	0.0145	7	0.4646	21

2.3 Fusion / Ensembling

We used the features/models described above to train six standard regression models: Linear Regression (Freedman, 2005), Ridge Regression (Hoerl and Kennard, 1970), Decision Tree Regression (Breiman et al., 1984), Random Forest Regression (Breiman, 2001), Gradient Boosting Regression (Friedman, 2000), and MLP Regression (Rosenblatt, 1958), all with default scikit-learn (Pedregosa et al., 2011) hyper-parameters. The models were trained on our training data split of the Google-MQM public dataset (see Section 2.1.1). For both the Phrase-only ensembles and the Collaboration ensembles we compared using all features as well as a "slick" version which uses only the best feature of each type: *pqps-xlmr-bal*, *palqa-wmt25* and *cuni-mr6-overall*. See results in Table 3. The submitted systems are marked with an asterisk. We selected the Gradient Boosting Regressor because it seems to achieve a good balance across both testsets as well as both metrics.

For a language-pair specific breakdown of results for the submitted systems refer to Table 7.

3 Task 2 (Phrase)

In Task 2, we experimented with using our internal AutoLQA systems that are based on LLMs and in-context learning. AutoLQA systems have been developed to produce the complete MQM annotation, i.e. they include the error category by default.

Table 3: Task 1 Phrase-only and Collaboration ensembles for different regression models and feature sets. Note on the submitted systems: phrase-slick is Phrase's primary submission, collab-slick is Phrase's secondary submission, collab-full is CUNI's secondary submission

Feature Set	Model	google-mqm		wmt22-qe	
		Kendall's τ	Pearson r	Kendall's τ	Pearson r
phrase-full	Linear Regression	0.2829	0.5517	0.3133	0.5092
15 features	Ridge Regression	0.2829	0.5517	0.3137	0.5095
	Decision Tree Regressor	0.2720	0.4976	**0.3533**	0.5047
	Random Forest Regressor	**0.3292**	0.5455	0.3006	0.4936
	Gradient Boosting Regressor	0.2912	0.5531	0.3423	**0.5250**
	MLP Regressor	0.2970	**0.5641**	0.3214	0.5115
phrase-slick	Linear Regression	0.2902	**0.5562**	0.3423	**0.5153**
2 features	Ridge Regression	0.2902	**0.5562**	0.3423	**0.5153**
	Decision Tree Regressor	**0.2962**	0.5431	0.3451	0.5052
	Random Forest Regressor	0.2706	0.5175	0.2937	0.4459
	Gradient Boosting Regressor**	0.2931	0.5549	**0.3474**	0.5145
	MLP Regressor	0.2921	0.5543	0.3444	0.5118
collab-full	Linear Regression	0.2844	0.5529	0.3179	0.5113
23 features	Ridge Regression	0.2842	0.5526	0.3184	0.5115
	Decision Tree Regressor	0.2720	0.4976	**0.3533**	0.5047
	Random Forest Regressor	**0.3327**	0.5553	0.2990	0.4940
	Gradient Boosting Regressor*	0.2931	**0.5558**	0.3455	**0.5302**
	MLP Regressor	0.2732	0.5369	0.3249	0.5145
collab-slick	Linear Regression	0.2934	0.5580	0.3424	0.5178
3 features	Ridge Regression	0.2934	**0.5581**	0.3424	0.5178
	Decision Tree Regressor	0.2930	0.5397	0.3507	0.5099
	Random Forest Regressor	0.2824	0.5209	0.3096	0.4660
	Gradient Boosting Regressor*	**0.2939**	0.5524	**0.3558**	**0.5213**
	MLP Regressor	0.2812	0.5484	0.3365	0.5108

Table 4: Task 2 F1 score for various engines. The results are reported on our Google-MQM test data. Note the `gpt-4o-mini-FT` is Phrase's fine-tuned version of the `gpt-4o-mini` model.

Engine	F1
`gpt-4o-mini`	0.1835
`gpt-4.1-mini*`	0.2277
`gpt-4o-mini-FT**`	0.2669

We used the Google-MQM test set as our development dataset, as described in Section 2.1.1.

We used GPT engines for this task. We experimented with both the off-the-shelf GPT models, as well as our fine-tuned (FT) versions of some of the models. Fine-tuning was performed on ∼100k segments of Phrase internal data. Table 4 depicts the performance increase when FT is used.

We also experimented with the wording mainly reducing the prompt from the complete MQM task to ESA, i.e., stripping off the error category part of the prompt. We are not allowed to disclose the full prompt but we have experienced that this reduction did not change the performance on our dev set. We have therefore decided to use one of our AutoLQA systems based on the `gpt-4o-mini-FT` model as our primary submission (denoted by double asterisk in Table 4).

Table 5: Task 3 Evaluation results for different systems. C-QE-Δ denotes the COMET-QE delta.

System	C-QE-Δ	TER	GtE-ratio
prod*	0.005	44.51	0.0001
onlyerrors	0.026	25.24	0.0010
accuracy	0.024	15.23	0.0016
fluency	0.050	37.05	0.0014
fluency_s**	0.043	28.27	0.0015

Table 6: Task 3 Primary system quality estimation scores. C-QE denotes the absolute COMET-QE score.

Dataset	baseline		post-edit	
	C-QE	QPS	C-QE	QPS
devtest	0.287	0.901	0.330	0.921
eval-full	0.055	0.878	0.073	0.901
eval-en-cs	0.079	0.910	0.102	0.923

Our secondary submission (denoted by *) is based on the FT gpt-4.1-mini. Due to higher cost, we have not evaluated the model on our development set and have relied on the results in Table 4 and on internal observations at Phrase where in most cases gpt-4.1-mini outperforms gpt-4o-mini.

4 Task 3 (Phrase)

For Task 3, our submissions are based on GPT models as well. We experimented with varying prompts, OpenAI models and settings (notably the amount of reasoning effort which can be set to "low", "medium" or "high" in the API).

After visual inspection and analysis of the error spans and scores provided for the development set, we deemed them quite noisy and not informative enough. We therefore opted for not using this information in the experiments. Our systems only utilise the source sentences and the MT outputs.

Our secondary submission prod (denoted by a single asterisk in Table 5) is loosely inspired by Phrase's production systems for automated MT adaptation. We submitted it as a baseline; however, this system is tailored to addressing typical customer use cases, such as correcting terminology, formality, or the placement of inline tags. Therefore, most of its instructions are not applicable to the WMT setting, so the system wastes its attention on irrelevant aspects. In addition, it required us to set a specific formality and without further domain-specific adjustments, the system likely changed the tone in many cases, leading to substantial amounts of post-editing.

We also experimented with dedicated prompts and a simpler setup targeted towards only improving the general quality of the translations. We simplified the system as well—whereas our production systems typically perform multiple passes over the data, here, we limited the system to just a single pass.

In terms of OpenAI models, we find that reasoning models are well suited for this task. We evaluated several combinations of reasoning effort settings and model types (notably o3 and o3-mini). Our primary submission, as well as most of the contrastive runs, utilise the o3[6] model with medium reasoning effort.

We evaluated four different prompts:[7]

- onlyerrors — focused on finding errors in the translation
- accuracy — improving translation accuracy
- fluency — improving translation fluency
- fluency_s — improving translation fluency, the prompt contains individual steps that the system should follow.

Based on the results on the WMT development set, we selected the prompt fluency_s as our primary submission. This system has the second-highest improvement in COMET-QE score and makes fewer edits than fluency, thereby reaching a better GtE-ratio (gain-to-edit ratio, i.e., the difference in COMET[8] divided by TER, as defined in the task description). It seems noteworthy that COMET-QE score grows more when the prompt is focused on improving translation fluency, as opposed to accuracy.

4.1 Analysis

Based on the official leaderboard, none of our systems improved over the baseline translations on the current test set. In order to explain this negative result, we carried out a post-submission analysis using the current evaluation data.

We found that based on COMET-QE, our system does improve the general translation quality.

[6]The full model name is o3-2025-04-16.
[7]We describe each in more detail in Appendix C.
[8]Approximated here using COMET-QE.

However, we also found COMET-QE to be quite unstable (see the absolute scores across different datasets in Table 6) and decided to complement it with our proprietary QPS metric (see Sec. 2.1.2 for reference). This again showed a consistent improvement in quality.

Finally, we carried out a blinded human evaluation using a small random sample of 50 English-Czech evaluation examples. We removed examples with very low edit rates (TER<10) prior to the sampling in order to prevent the annotator from spending time comparing very similar outputs. The annotator saw each candidate translation side by side (ordered randomly; one being the original MT output and the other our post-edited version) and rated each output on the scale of 1 to 10.

Consistently with the automated quality estimation results (COMET-QE and QPS), the annotator rated the post-edited outputs as higher quality. Out of the 50 examples, our system output was preferred in 31 cases and tied in 10, whereas the baseline output was only preferred in 9 instances. The average score was 8.48±1.20 and 7.50±1.53 (approximate randomization test p-value≈0.002) for the system and baseline respectively, illustrating that while translation quality was quite high overall, translations were not generally deemed perfect by the annotator.

These findings leave conclusions quite open; on the one hand, both automated QE metrics and the human annotator (albeit on just a single language pair) showed clear preference for our post-edits. On the other hand, the official evaluation (which uses reference-based COMET) does not rate the system as better than the baseline. Further analysis leveraging also the official reference translations would be needed to explain this apparent contradiction.

5 Conclusion

We presented our submissions to WMT25 MT Evaluation Task developed by two teams, CUNI and Phrase: two primary and four secondary submissions for Task 1, one primary and two secondary submissions for Task 2 and one primary and one secondary submission for Task 3.

For system fusion/ensembling used in Task 1, the rule "more features are better" generally holds, although not for every individual language-pair or metric. It turns out that the ability of the regressor model "saturates" and gets stuck in the diminishing-reward paradigm soon.

It is interesting to see that competitive results can be achieved by combining very few systems if we choose them well, as is the case of "slick" systems.

In Task 3, we failed to improve on the baseline translation quality in the official rankings. However, our internal post-submission analysis showed a different picture and we currently lack a good explanation for the disagreement.

During the submission, we saw few puzzling outliers for certain language pairs in all tasks' leaderboard, and we expect to reveal the issues after the official annotations are released.

6 Acknowledgment

This work has received funding from the Project OP JAK Mezisektorová spolupráce Nr. CZ.02.01.01/00/23_020/0008518 named "Jazykověda, umělá inteligence a jazykové a řečové technologie: od výzkumu k aplikacím."

References

L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. 1984. Classification and regression trees.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

David Freedman. 2005. *Statistical Models : Theory and Practice*. Cambridge University Press.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of

human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jerome H. Friedman. 2000. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. DSPy: Compiling declarative language model calls into self-improving pipelines.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. 2024. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9340–9366, Miami, Florida, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

F. Rosenblatt. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Ole Tange. 2025. Gnu parallel 20250622. GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, et al. 2024. mgte: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. *arXiv preprint arXiv:2306.07899*.

A Task 1 Language-Pair Specific Breakdown

Table 7: Task 1: Language-pair specific breakdown for the submitted systems

		google-mqm		wmt22-qe	
Submission	LP	Kendall's τ	Pearson r	Kendall's τ	Pearson r
phrase-slick	en-de	0.3282	0.6240	0.3089	0.4660
	en-es	0.3293	0.6466	—	—
	en-ru	0.4037	0.6408	0.3571	0.4885
	en-zh	0.2602	0.5467	—	—
	he-en	0.3099	0.4841	—	—
	ja-zh	0.1878	0.3163	—	—
	zh-en	0.2767	0.5971	0.2184	0.3463
palqa-wmt25	en-de	0.2867	0.4886	0.3005	0.3962
	en-es	0.3123	0.5699	—	—
	en-ru	0.3323	0.4341	0.3525	0.4405
	en-zh	0.2105	0.4201	—	—
	he-en	0.3300	0.4829	—	—
	ja-zh	0.2460	0.4050	—	—
	zh-en	0.2502	0.4458	0.2180	0.2636
collab-slick	en-de	0.3232	0.6156	0.3307	0.4805
	en-es	0.3337	0.6417	—	—
	en-ru	0.4016	0.6394	0.3783	0.5011
	en-zh	0.2532	0.5398	—	—
	he-en	0.3189	0.4903	—	—
	ja-zh	0.1964	0.3016	—	—
	zh-en	0.2843	0.6013	0.2207	0.3493
cuni-mr6-overall	en-de	0.1608	0.3648	0.3736	0.5124
	en-es	0.2667	0.4966	—	—
	en-ru	0.2549	0.4706	0.3850	0.4754
	en-zh	0.1142	0.3932	—	—
	he-en	0.2746	0.4372	—	—
	ja-zh	0.2220	0.2402	—	—
	zh-en	0.1632	0.3069	0.1540	0.1889
collab-full	en-de	0.3638	0.6730	0.3235	0.4525
	en-es	0.3486	0.6209	—	—
	en-ru	0.4562	0.7262	0.3790	0.5264
	en-zh	0.2471	0.5099	—	—
	he-en	0.2773	0.4692	—	—
	ja-zh	0.1614	0.2578	—	—
	zh-en	0.3371	0.6908	0.1906	0.3614

B Task 1 - DSPy Models Optimized Prompts

DSPy's resulting "programs" for our *mr** submissions, based on which the final single prompt requesting all the scales or the individual prompts are constructed. These include the selected n-shot examples and any adjustments to the initial data field descriptions or instructions.

B.1 mr6

The full version of optimised "program" (instructions and selected few-shot examples) for *mr6* can be found in the GitHub repository.[9]

B.2 mr7.2.1

The full version of optimised "program" (instructions and selected few-shot examples) for *mr7.2.1* can be found in the GitHub repository.[10]

C Task 3 Prompts

In this section, we share the details of instructions specifically tailored for our WMT systems.

There is also a common core part of the prompt which sets up the task and specifies processing requirements. This part is proprietary and therefore not shared here.

- `onlyerrors` Focus only on errors (grammar, fluency, mistranslation etc.), do not perform subjective or preferential edits.
- `fluency` Concretely, the task is to improve the overall fluency and naturalness (with translation accuracy being important but secondary), while minimising the number of edit operations. Avoid translationese and prefer loose translation when it allows for a more idiomatic translation.
- `fluency_s` Concretely, the task is to improve the overall fluency and naturalness (with translation accuracy being important but secondary), while minimising the number of edit operations. Avoid translationese and prefer loose translation when it allows for a more idiomatic translation.

  Let's go step by step:

  – Identify errors, translationese, or disfluent spans in the translation.
  – Consider several possible corrections.
  – Out of the possible outputs, choose one that is the most natural-sounding and requires few changes (edit operations).
- `accuracy` Concretely, the task is to improve the overall accuracy (with translation fluency being secondary), while minimising the number of edit operations.

[9] https://github.com/hrabalm/wmt25-mt-eval-task/blob/main/mr6/best.json
[10] https://github.com/hrabalm/wmt25-mt-eval-task/blob/main/mr7.2.1/best.json

MSLC25: Metric Performance on Low-Quality Machine Translation, Empty Strings, and Language Variants

Rebecca Knowles **Samuel Larkin** **Chi-kiu Lo** 羅致翹
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
{rebecca.knowles,samuel.larkin,chikiu.lo}@nrc-cnrc.gc.ca

Abstract

In this challenge set, we examine how automatic metrics for machine translation perform on a wide variety of machine translation output, covering a wider range of quality than the WMT submissions. We also explore metric results on specific types of corner cases, such as empty strings, wrong- or mixed-language text, and more. We primarily focus on Japanese–Chinese data, with some work on English and Czech.

1 Introduction

This paper describes a challenge set submitted to the Challenge Set Subtask of the (Unified) Machine Translation Evaluation shared task at the 2025 Conference on Machine Translation (WMT); we focus primarily on segment-level quality score prediction with a brief preliminary note on word-level error detection and span annotation. For this third iteration of the Metric Score Landscape Challenge (MSLC),[1] we run a smaller set of experiments. We once again focus on Japanese→Chinese news translation for the task of exploring the range of low- to mid-quality MT (as compared to the high-quality MT systems submitted to WMT). We also include analysis of empty strings (as in past iterations), mixed- and wrong-language text, and a preliminary note on English spelling variation; for these tasks, we use a mix of Japanese→Chinese data, English language data, and Czech→English data. Our approach, particularly for low- to mid-quality MT and for empty strings, demonstrates a low-cost way to test metrics on the wider quality landscape. We encourage developers of metrics to run such evaluations themselves prior to releasing metrics. For developers of metrics who are unable to run such evaluations themselves, we call on them to explicitly declare that their released metrics have not been tested on low- or mid-quality MT and should not be used for such cases without additional testing.

[1] MSLC data and additional figures can be found at https://github.com/nrc-cnrc/MSLC.

2 Prior and Related Work

This work describes the third MSLC challenge set. In Lo et al. (2023b), the first iteration, our intent was to focus primarily on low- and mid-quality MT output across four language pairs. The process of evaluation that year brought to light two other issues: the scores that metrics assigned to empty strings (as some MT system submission that year included empty strings, providing a natural experimental set) and "universal scores" (scores that were assigned very frequently by certain metrics) similar to the "universal translations" described in Yan et al. (2023). The second iteration of MSLC, Knowles et al. (2024), continued the work on low- and mid-quality MT, with the addition of experiments on empty strings, mixed- and wrong-language text, and language variants (in that instance, Spanish language terminology that differs across language variants). In concurrent work to MSLC24, Zouhar et al. (2024) proposed COMET-specific mitigations to many of the issues observed in both papers: incorporating language ID in order to mitigate issues with mixed- or wrong-language text, using signatures in the spirit of `sacreBLEU` (Post, 2018) to help explain variations in metric output, and the issue of empty strings.

There is a tradition of challenge sets targeting specific linguistic phenomena for MT (Isabelle et al., 2017; Burlot and Yvon, 2017; Guillou et al., 2018; Rios et al., 2018; Stanovsky et al., 2019, i.a.), while challenge sets for evaluation (i.e., challenge sets targeted at metrics) tend to be relatively newer (see, i.a., the descriptions of the challenge sets at the Metrics shared tasks: Freitag et al., 2022, 2023, 2024). Of these challenge sets, Amrhein et al. (2022, 2023) also explore wrong-language text (among 68 phenomena across a large number of language pairs in the ACES challenge set),

Proceedings of the Tenth Conference on Machine Translation, pages 945–956
November 8-9, 2025

noting particular issues for reference-free metrics, similar to our observations.

Our work is situated more broadly in the area of MT evaluation and corner cases for MT evaluation. While there is prior work focusing on specific metrics and corner cases (Hanna and Bojar, 2021; Amrhein and Sennrich, 2022; Yan et al., 2023; Zouhar et al., 2024, i.a.), submitting this challenge set to the shared task permits us to examine and compare performance on corner cases and MT quality ranges across metrics in a controlled environment. Importantly, readers should note that this paper is a description of a challenge set submitted to a shared task, rather than a complete, in-depth exploration of all the areas on which it touches. We note, in both the limitations section and throughout the work, that there are some components that represent established evaluations (MSLC-A and the empty strings work) while other components are presented as initial proof-of-concept experiments to determine whether future in-depth evaluation is indicated (these smaller preliminary experiments should not be used to draw sweeping conclusions).

3 Data

The MSLC challenge set is divided into two main components: MSLC-A (which covers low-quality and mid-quality MT) and MSLC-B (which targets specific corner cases and potential challenges for metrics).

3.1 MSLC-A

The MSLC-A portion of the challenge set focuses on covering a range of MT quality, from extremely low quality (incomprehensible output) to mid-quality output. The intention is to fill some of the gaps in evaluation of new metrics, which are typically tested at WMT on high-quality systems only, despite the fact that they may go on to be used for a wider range of quality in practice.

We use Japanese→Chinese systems from Knowles et al. (2024). The MT models were all constrained (as per the 2024 WMT General Task rules) NMT models built using Sockeye version `3.1.31` (Hieber et al., 2022) and trained on WMT training data; for a more detailed description of the systems, see Larkin et al. (2024). We translate and use only the News portion of the 2025 WMT General Task data.

There are six MT systems in this part of the challenge set. The lowest-quality system is indicated with the letter A, and the quality approximately increases as the system labels proceed alphabetically. Using the same process described in Larkin et al. (2024) and Knowles et al. (2024), these low- to mid-quality outputs were produced by translating the same source text[2] using early checkpoints saved during model training, with the lowest quality produced by the early checkpoints (i.e., when the system produced nonsensical and repetitive output) and the mid-quality outputs produced by later checkpoints. These low- to mid-quality MT system outputs were ranked by *BLEU* and manually examined (on a subset of the data) by an author fluent in the target language to confirm their increasing quality.[3] We run a limited version of the MSLC-A experiments in this edition of the challenge set, without submission of the systems to the General Task at WMT (i.e., we do not have human evaluations that indicate the magnitude of the gap, if any, between our highest-performing mid-quality MT system and the lowest-performing submitted system).

3.2 MSLC-B

For the MSLC-B portion of the challenge set, we focus on three different types of edge cases for metrics: empty strings, mixed- and wrong-language text, and English spelling variants.

In past iterations of MSLC, we observed that differences in what a metric treats as a "document" can have an impact on the scores it assigns in our test sets. In an effort to ensure that document-level metrics did not mix together the contrastive examples we were having scored, we appended strings to the document IDs to identify these as "separate documents" where appropriate (i.e., when two contrastive examples were from the same document, they might be assigned as "[DOCID]-1" and "[DOCID]-2" so that the metric should treat them as separate documents).

3.2.1 Empty Strings

Given past observations (Lo et al., 2023b; Knowles et al., 2024) of unusual outputs related to empty

[2]Text was translated at the segment or sentence level and then re-collected into documents.

[3]While it may be desirable to perform more formal evaluation, the lowest-quality systems are of such low quality as to be visibly "nonsense" even to non-speakers of the language as well, with a clear trend of improvement. Thus, while we do not have MQM or other manual scores to rank these, we can be confident in the overall trend, and particularly confident that, e.g., system A is a substantially worse system than system D.

strings, such as surprisingly high scores, we once again examine this topic. We use a small Japanese→Chinese dataset for this: 10 punctuation characters, 10 words, 10 phrases, and 10 segments (sentences or larger); all except for the punctuation are selected from the news domain portion of the WMT 2025 General Task test data. For each of these, we explore the case where we have an empty source and reference and a non-empty hypothesis (representing overgeneration: an MT system producing something from nothing) and the case where there is an empty hypothesis with a full source and reference (undergeneration: an MT system producing nothing when it should have produced something).

3.2.2 Mixed- and Wrong-Language Text

For metrics that rely on multilingual embeddings and ones that do not take into account the intended source and target language, there is a risk of returning high scores for wrong-language output. As in Knowles et al. (2024), we examine metric scores when presented with wrong-language or mixed-language hypotheses. In this case, we take advantage of the multi-way parallel test sets (pivoted on English) by using segments from English news test data with its Chinese and Japanese translations. We run these experiments on 18 segments with Japanese→Chinese as the intended language pair and translation direction, using the Japanese data as the source, the Chinese data as the reference, the English data as the wrong-language hypothesis, and a pseudo-codeswitched mix of English and Chinese as the mixed-language hypothesis.[4]

3.2.3 English Language Spelling Variants

While recent iterations of WMT have included more regional specifications regarding language variants, variations in English have been overlooked. We use a word list of common UK (en_GB) and US English (en_US) spelling differences[5] to select segments from the Czech→English news test data that contain words with the potential for different spellings.[6] We then automatically produce two versions of each of these 20 English segments, one with standard UK spellings and one with standard US spellings, and confirm them with manual examination. The English sentences are otherwise identical: minimal pairs where the only difference is the spelling of the words of interest. We then inverted the translation direction, treating Czech as the source and English as the target. The official submission format for the Metrics Challenge Set subtask included a field for language ID for the target; we submitted versions that included the region (en_US or en_UK) and one that did not (en). This should enable us to see whether some metrics utilize this information and, for those that do not, whether there is a bias towards a particular language variant. Of note, this is most relevant to reference-free metrics, as the reference is either an exact match to the hypothesis or identical except for the spelling of the words of interest. This is a small preliminary experiment to determine whether future large-scale evaluation of this topic may be fruitful.

4 Metrics

We focus on analyzing the scores produced by the baseline metrics and the primary submissions. There are 9 baseline metrics (including 2 "sentinel" metrics designed to scrutinize the metric meta-evaluation process) and 5 primary metrics that participated in portions of our challenge set for segment-level evaluation. One baseline and 2 primary metrics participated in the error span detection portion of our challenge set.

The segment-level score baselines are *BLEU* (Papineni et al., 2002), *spBLEU* (NLLB Team et al., 2022), *chrF* (Popović, 2015), *BERTScore* (Zhang et al., 2020), *COMET-22* (Rei et al., 2022a), *CometKiwi* (Rei et al., 2022b), *YiSi-1* (Lo, 2019), *sentinel-cand* and *sentinel-src* (Perrella et al., 2024).

For the segment-level quality score prediction task, five metrics participated in our experiments. *MetricX-25* (Juraska et al., 2025), an updated version of *MetricX*, is an encoder-only regression model initialized from Gemma 3 (Team et al., 2025) 12B and fine-tuned on publicly available DA and MQM scores from WMT 2015–23. *mr7.2.1* (Hrabal et al., 2025) is based on the Gemma 3 27B IT model and is prompted with the DSPy (Khattab et al., 2024) framework and its MIPROv2 opti-

[4]This mixed-language data was produced manually by an author fluent in the languages and is intended to contain the full semantic content of the text in such a way that it could be read by someone who speaks both languages and be perceived as similar to naturally generated by code-switching speakers.

[5]https://github.com/hyperreality/American-British-English-Translator/blob/master/data/american_spellings.json

[6]In some cases, we shortened the segments to more tightly focus on sentences containing the words of interest (shortening both the English and Czech sides of the sentence pair).

mizer. *Polycand-2* (Züfle et al., 2025) is a COMET-based metric that incorporates two alternative translations of the same source segment (provided by other translation systems) to better contextualize and assess the quality of the translation being scored. *rankedCOMET* (Maharjan and Shrestha, 2025) is a COMET-based metric post-processed with rank normalization for each language pair. *UvA-MT* (Wu and Monz, 2025) calibrates quality estimation and likelihood on the Gemma 3 12B IT model, then directly uses the token average likelihood as a metric for quality estimation.

For the error span detection task, three systems participated in our MSLC-B experiments: baseline *XCOMET* (Guerreiro et al., 2024) and the two primary submissions of *AIP1* (Yeom et al., 2025) and *GemSpanEval* (Juraska et al., 2025). *AIP1* uses the OpenAI o3 (OpenAI, 2025) reasoning model and its structured-output mode to detect translation errors at the span level. *GemSpanEval* is based on the Gemma 3 27B model and is finetuned to predict MQM error spans.

Some metrics use reference translations in the process of producing their scores, while others do not. Throughout the remainder of the paper we indicate the 3 baseline and 4 primary reference-free (QE) metrics—those that do not use the reference translation in their scoring—with an asterisk before the metric name. There are 3 reference-free primary metrics (**mr7.2.1*, **Polycand-2*, **UvA-MT*) in the segment-level score prediction task and 1 (**AIP1*) in the error span prediction task.

5 MSLC-A Results and Plots

Interpretation note of caution: our submitted MSLC-A challenge set was scored at the document level by the automatic metrics, while the submitted primary MT systems from the general task were scored at a sub-document segment level by the automatic metrics. In order to be able to compare these, we have averaged the segment-level scores to produce a document-level score for each document. We note that this may not always be identical to the score that the metric would have assigned had it scored the full document directly, and caution should therefore be taken when drawing conclusions.

Figure 1 shows system average scores for the MSLC-A systems (cool colours, left) and the systems submitted the the WMT General MT task, computed only over the News data. These averages are computed from the document-level scores, which in the case of MSLC-A data were produced directly by the metrics and which in the case of the submitted systems were produced as an average of segment-level scores. We find several points of interest. Three of the metrics, **COMETKiwi22*, *MetricX-25*, and **UvA-MT* have some difficulty with the rankings of the lowest-quality MSLC-A systems. In particular, **UvA-MT* ranks the worst system (whose output is almost entirely nonsensical and unreadable) as similar in quality to the high-quality WMT submissions. This indicates that these metrics—for this language pair at least—may not be trustworthy metrics to use when trying to evaluate low-quality MT. Metric users should consider alternative choices of metrics if they have reason to believe that their MT output may be of low quality or mid-range quality. This could also have an impact on MT systems trained using these metrics, particularly in the early stages of training.

We also observe that some metrics (**Polycand-2*, *COMET22*, i.a.) devote a very small portion of their metric's space of possible scores to the high-quality systems, with a wider range of scores for the low-quality systems, while others like *chrF* distribute the score range more evenly. This may impact how useful a metric is for distinguishing between systems of different levels of quality, and may also play a role in human interpretation of metric score differences; for broader discussion of metric score differences and human interpretations thereof, see Mathur et al. (2020) and Lo et al. (2023a), i.a. There is not an inherent right answer to how a metric should use its score space, rather, it is tied to the intended use of the metric.

We observe that some metrics show overlap or near overlap between the best of the MSLC systems and the lowest-scoring of the submitted systems while others show a large gap between them. Since we do not have full human annotations available, we cannot make any claims about whether there would be an overlap or a gap based on human evaluation.

Figure 2 provides another way to visualize the metric scores, showing histograms of scores assigned to each system along the diagonal, and scatterplots showing correlations between metrics on the off-diagonals. The metrics that give higher-than-expected scores to the low-quality systems once again stand out where they do not correlate with systems that rank the low-quality systems as

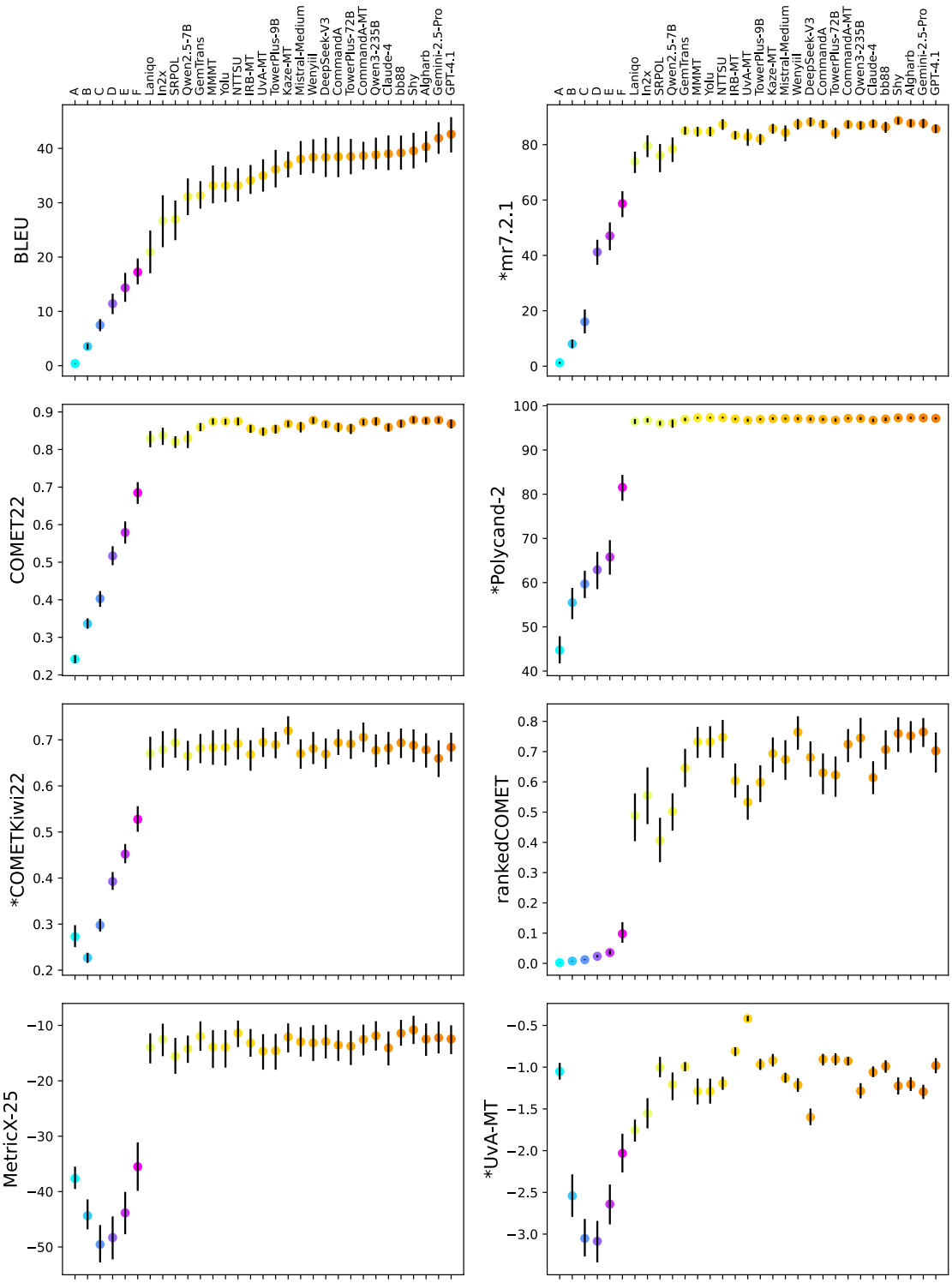


Figure 1: System average scores for Japanese→Chinese. MSLC systems (cool colours, left) are ordered by BLEU score and brief manual examination; WMT submitted systems are ranked by average BLEU score.

expected. We can also use the histograms to gain a better understanding of the score distributions assigned to the data, such as the very low scores and comparatively small score range assigned by *rankedCOMET* to the low-quality systems, as compared to other metrics, or the somewhat bimodal score distribution from **mr7.2.1*. In past iterations of MSLC, this type of figure was particularly useful in highlighting unusual properties of some of the metrics, such as discretizing the score space or

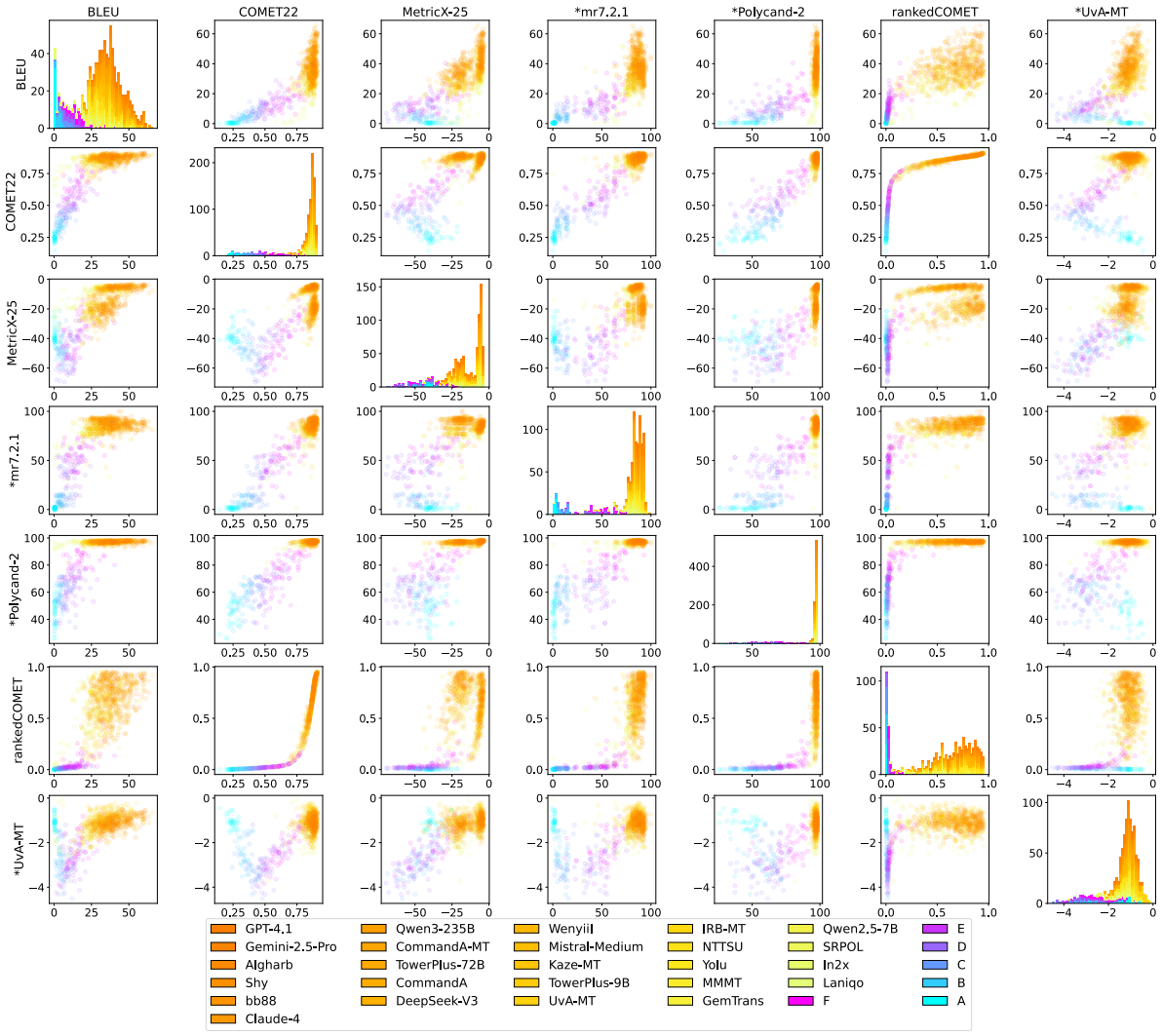


Figure 2: Matrix of segment-level scores for Japanese→Chinese. Along the diagonal are stacked histograms of segment scores across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top). The off-diagonal entries are scatterplots where each point is a single document positioned according to the score assigned to it by row and column metrics; each point is coloured according to the same colours as the histogram.

assigning specific scores very frequently ("universal scores"). We do not observe such results in this year's set of metrics.

Due to scheduling constraints, we did not receive the human annotation scores in time to display those in these figures; we plan to incorporate them into final additional figures on the MSLC website (https://github.com/nrc-cnrc/MSLC).

6 MSLC-B Results and Plots

6.1 Empty Strings

In Figure 3 we show the scores assigned by the five primary submission metrics to punctuation, words, phrases, and segments (sentences to documents) when those are paired with an empty source and reference. The vertical red lines indicate the minimum and maximum scores assigned by the same metric to all WMT General Task primary submissions on the News portion of the data; since different metrics use different score ranges, this is used to provide the reader with some context about where the scores for these corner cases fall in comparison to scores assigned to more usual MT output. These empty string examples are fairly extreme examples of MT failures; string-based metrics like *BLEU* would assign them scores of 0.

We see several types of responses. The metric **mr7.2.1* assigns its lowest score to all of these, similar to what we observe from metrics like *chrF* and *BLEU* (not shown in figure); this is arguably what we would expect, since an empty source should produce an empty hypothesis. The results from

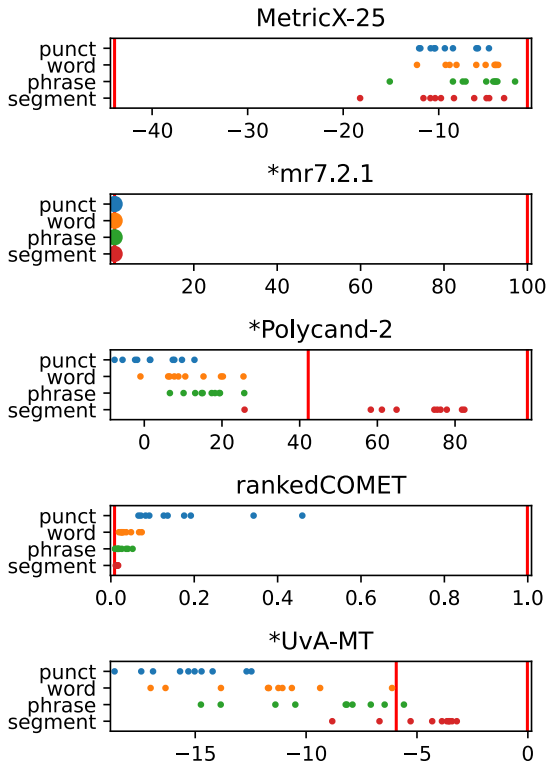


Figure 3: Japanese→Chinese scores assigned to text when paired with empty source and reference. Where multiple strings receive the same score, this is indicated by proportionally increased dot size (in the case of **mr7.2.1*, all strings received the same score of 0, as indicated by the large dots on top of the left red vertical line). Red vertical lines indicate the minimum and maximum scores assigned over all Japanese→Chinese WMT News primary submission data. Asterisks indicate reference-free (QE) metrics.

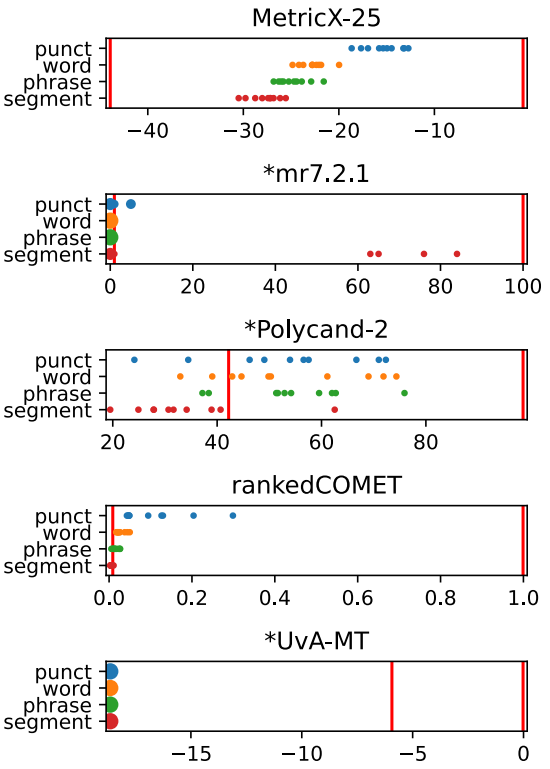


Figure 4: Japanese→Chinese scores assigned to empty string hypothesis paired with real (non-empty) source and reference. Where multiple strings receive the same score, this is indicated by proportionally increased dot size (in the case of **UvA-MT*, all strings received the same score, as indicated by the large dots). Red vertical lines indicate the minimum and maximum scores assigned over all Japanese→Chinese WMT News primary submission data. Asterisks indicate reference-free (QE) metrics.

rankedCOMET show a pattern that was observed in Knowles et al. (2024), where longer strings (segments) receive lower scores and some of the shorter strings (punctuation) receive higher scores, matching the intuition that a full sentence or document is more different from the empty string than a single character is. For both **Polycand-2* and **UvA-MT* we observe the opposite trend, with higher scores assigned to the longer strings. In both cases, most of the scores assigned to punctuation, words, and phrases are lower than the scores assigned to any of the submitted MT system data, but some of the scores assigned to the segments fall close to the middle of that score range. While both are reference-free metrics (i.e., not using the information that the reference string is the empty string), they do still have access to the source, making the result on segments more surprising.

The case of empty source and reference paired with non-empty hypothesis is an extreme representation of overgeneration, generating text that is not grounded in the source. We argue that unusual results on this set of data should raise questions for more exploration of a metric's performance on instances of overgeneration. It will require additional study to determine if there is a link between these, or if these results are confined to this particular corner case.

Figure 4 is the corresponding figure for the empty string hypothesis paired with a real (non-empty) source and reference. This is an extreme case of undergeneration (failing to generate any output). Two reference-free metrics, **mr7.2.1* and **UvA-MT*, assign very low scores to most of these examples, though **mr7.2.1* assigns scores in the top half of its score range to some segments. The results for *rankedCOMET* are quite similar to the results on the previous set of experiments, with slightly higher scores assigned to punctuation, but generally low scores overall. *MetricX-25* follows a similar pattern with the shorter strings scoring higher, but overall in the middle of the score range, while **Polycand-2* does not show a clear pattern.

Once again, we argue that assigning relatively high scores to empty string hypotheses may indicate that metrics are failing to pick up on undergeneration. Additionally, assigning non-lowest scores to the empty string presents a potential mismatch between metrics and typical standards for human evaluation (i.e., human annotators instructed to, or otherwise deciding on their own to, give the lowest scores to empty translations).

6.2 Mixed- and Wrong-Language Text

Metric	mix>wrong	equal	wrong>mix
**COMETKiwi22*	1	0	17
MetricX-25	2	0	16
**mr7.2.1*	18	0	0
**Polycand-2*	17	0	1
**UvA-MT*	1	0	17

Table 1: Comparison of scores of mixed and wrong-language text. Systems indicated with an asterisk (*) are reference-free (QE) metrics. All baseline reference-based metrics ranked all 18 mix>wrong.

As described in Section 3.2.2, we explore the scores that metrics assign to mixed- and wrong-language text. For our Japanese→Chinese challenge set for this task, the mixed language text is a mix of English and Chinese, intended to contain the full semantic information of the source. The wrong language text is English. Both the Chinese reference (used to build the basis of the mixed-language text) and the Japanese source are actually translations of the English data (which we also use to construct the mixed-language text). Due to the overlap between the Chinese reference and the mixed-language hypothesis, almost all baseline and primary reference-based metrics score the mixed-language hypothesis higher than the wrong-language for all 18 examples. The one exception to this is *MetricX-25*, as shown in Table 1, which scores the wrong-language text higher than the mixed-language text in 16 of the 18 examples. Both **COMETKiwi22* and **UvA-MT* score the wrong-language text above the mixed-language text in 17 out of 18 examples, while **Polycand-2* does the reverse and **mr7.2.1* prefers the mixed language text in all cases.

It remains an open question how mixed-language text should be scored, and is likely dependent on the intended audience of the translation. In any case, it may be surprising to observe systems preferring hypotheses that contain none of the intended target language at all over those that do at least include some target language text. This highlights—particularly with the shift to reference-free and multilingual metrics—the importance of taking into account the intended target language in evaluation. While we use a very small dataset here (18 segments), the consistency that we observe within metrics is notable. The issue of wrong-language output continues to be one that appears to be under-

examined by the designers of metrics, a claim we make based not only on this small-scale proof-of-concept, but by similar work in past challenge sets across more languages (Amrhein et al., 2022, 2023; Knowles et al., 2024).

6.3 English Language Spelling Variants

With the inclusion of an increasing amount of region information for the languages in WMT, we were interested in exploring English language spelling variants. As a preliminary step, we explored common British and American spelling differences. We used pairs of Czech→English segments where the English hypothesis varies only in the spelling conventions for certain terms. We submitted this portion of the challenge set three different times, once each with the language/region described as "en", "en_GB", and "en_US". We observed no difference in metric preferences depending on the choice of region descriptor; it is likely that most of these metrics are not taking into account the regional information at this granularity (compare also to the results in Section 6.2, which suggest that even the language code itself may not be entirely influential). For the three reference-free metrics that participated in this portion of the challenge set, we observed three different results (Table 2): **COMETKiwi22* was equally split between US and GB, but rarely scored them identically, **mr7.2.1* scored them identically more than half of the time and preferred GB almost half of the time, and **Polycand-2* scored the US variant higher the majority of the time. Due to the setup of the experiment, we could also check whether repeated instances of the same examples were scored identically; for **Polycand-2* there were some small (up to $3.81e-06$) differences in repeated scores; the differences between the US and GB variants were substantially larger.

We manually examined the error span results for baseline *XCOMET* and the two primary submissions of **AIP1* and *GemSpanEval* but did not observe any clear patterns related to the spelling variants. *GemSpanEval* did label some of the spelling variant terms as errors, but there was not a clear pattern related to the intended target language variants.

As the WMT shared tasks shift to include more region information, we expect that metrics will seek to handle this as well. We choose English for this particular example, because variation in English has been overlooked at WMT, even in instances when Englishes are paired with regionally-specified language variants. While we focused on English variation in the target; metric biases may also be relevant where the source is concerned.

Metric	US>GB	Equal	GB>US
**COMETKiwi22*	9	2	9
**mr7.2.1*	2	11	7
**Polycand-2*	16	0	4

Table 2: Comparison of reference-free (QE) metrics on pairs of Czech→English sentences where the English hypothesis only varies in whether certain terms use British or American English spelling conventions. For the 20 examples, the table shows the counts of those for which the US spelling version was given a higher score, for which the scores were equal, and for which the British spelling convention was given a higher score. Results are identical regardless of whether the intended target language has the region specified or not ("en", "en_GB", "en_US").

Both the dataset we used and the number of metrics that completed the task are much too small to draw broader conclusions from. Nevertheless, we think this will be an interesting avenue to explore, as WMT shifts to incorporate more regional information into its translation tasks.

7 Conclusions

We observe similar results to past MSLC experiments, with some metrics struggling to accurately score extremely low-quality (nonsensical) MT outputs. We continue to encourage discussion around how metrics should score empty strings and encourage additional analysis of how this does or does not correlate with broader metric sensitivity to over-generation and undergeneration. As we see more metrics shifting to use multilingual embeddings our large language models and more reference-free metrics, we encourage metric builders to consider how to incorporate information about the intended target language into their metrics (see also, Zouhar et al. (2024)). While it may be easy for a human—even one who cannot read the languages in question—to tell if an MT system has erroneously generated English when Chinese was expected, it may not be so simple for more similar language pairs. We would encourage metric builders to consider how to incorporate intended target language into their systems, and note that this may be an area where ignoring available references may have a real cost when it comes to metric trustworthiness.

Limitations

We focus on a small set of language pairs (in fact, smaller than in past iterations) and use small dataset sizes. This year, we did not submit systems to the General MT task, which means that we do not have a way to confirm how close those systems are (by human evaluation) to submitted systems; this may result in a gap in coverage between low and high quality systems. In general, these experiments represent corner cases that metrics builders should be considering in their systems. Our results primarily serve to flag issues to potential users of metrics and to encourage builders of metrics to test their metrics extensively.

Acknowledgements

We thank the WMT Metrics Task and WMT General MT Task organizers for permitting us access to the references in order to build this challenge set. We thank the reviewers for their comments and suggestions.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712, Singapore. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv*, abs/2207.05851.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. Cuni and phrase at wmt25 mt evaluation task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2024. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, pages 475–491, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Larkin, Chi-Kiu Lo, and Rebecca Knowles. 2024. MSLC24 submissions to the general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 139–146, Miami, Florida, USA. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023a. Beyond correlation: Making sense of the score differences of new MT evaluation metrics. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 186–199, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023b. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799, Singapore. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. Rankedcomet: Elevating a 2022 baseline to a top-5 finish in the wmt 2025 qe task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2025. Openai o3 and o4-mini system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine*

Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Di Wu and Christof Monz. 2025. Uva-mt at wmt25 evaluation task: Llm uncertainty as a proxy for translation quality. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and JinYeong Bak. 2025. Tagged span annotation for reasoning llm-based translation error span detection. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. Comet-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

MetricX-25 and GemSpanEval: Google Translate Submissions to the WMT25 Evaluation Shared Task

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Dan Deutsch, Pidong Wang, and **Markus Freitag**
Google Translate
{jjuraska,domhant,freitag}@google.com

Abstract

In this paper, we present our submissions to the unified WMT25 Translation Evaluation Shared Task. For the Quality Score Prediction subtask, we create a new generation of MetricX with improvements in the input format and the training protocol, while for the Error Span Detection subtask we develop a new model, GemSpanEval, trained to predict error spans along with their severities and categories. Both systems are based on the state-of-the-art multilingual open-weights model Gemma 3, fine-tuned on publicly available WMT data. We demonstrate that MetricX-25, adapting Gemma 3 to an encoder-only architecture with a regression head on top, can be trained to effectively predict both MQM and ESA quality scores, and significantly outperforms its predecessor. Our decoder-only GemSpanEval model, on the other hand, we show to be competitive in error span detection with XCOMET, a strong encoder-only sequence-tagging baseline. With error span detection formulated as a generative task, we instruct the model to also output the context for each predicted error span, thus ensuring that error spans are identified unambiguously.

1 Introduction

Large language models (LLMs) have been evolving at a breakneck speed over the past couple of years, achieving performance in machine translation (MT) that matches or even exceeds that of humans for certain languages and domains (Kocmi et al., 2024a). While human evaluation is still the most reliable way to assess the quality of MT models, in this fast-paced environment of state-of-the-art models being released on a monthly basis, the cost and duration make it infeasible to run human evaluation studies to benchmark models regularly. Improving automatic metrics is therefore instrumental to making further progress in MT, especially as the field expands to low-resource languages and more difficult domains.

Currently, the most successful automatic MT evaluation metrics are trained neural models themselves, following one of two main paradigms: (1) regression models predicting a scalar quality score (Rei et al., 2022a; Juraska et al., 2024), and (2) sequence-tagging or generative models, providing fine-grained quality feedback, including error spans, severities and categories (Fernandes et al., 2023; Guerreiro et al., 2024). In this work, we push the performance of automatic metrics in both categories higher by leveraging a state-of-the-art multilingual open-weights LLM, Gemma 3 (Gemma Team et al., 2025), resulting in two separate submissions to the WMT25 Evaluation Shared Task.

For the Quality Score Prediction subtask, we develop *MetricX-25*, a successor to MetricX-24 (Juraska et al., 2024), updated to use an encoder-only architecture and trained on a combination of publicly available direct assessment (DA) and Multidimensional Quality Metrics (MQM) scores from WMT shared tasks between 2015 and 2023. For the Error Span Detection subtask, we introduce *GemSpanEval*, a generative model that identifies and categorizes error spans in JSON format, trained exclusively on MQM error span annotations from WMT20–24. We enable GemSpanEval to uniquely identify short, non-unique error spans by training the model to also indicate the error span context where necessary.

The key takeaways from our experiments, detailed in this report, include:

1. Gemma 3 offers a strong multilingual foundation for an automatic MT evaluation metric, and adapting it to an encoder-only architecture proves highly effective for score prediction.
2. It is possible to train an automatic metric to effectively predict different types of score using a single regression head by simply mixing training examples of different scores, with a score type indication included in the input.

Proceedings of the Tenth Conference on Machine Translation, pages 957–968
November 8-9, 2025

3. Fine-tuning a strong multilingual decoder-only model can be competitive with encoder-only error span detection models.
4. Predicting error span context can be used to uniquely identify error spans when formulating span detection as a generative task.

2 Data

The systems we developed for both the quality score prediction and the error span detection are trained solely on publicly available data from the WMT Metrics shared tasks between 2015 and 2024. Quality ratings for system translations across those 10 years were collected using 3 different methods:

1. Direct assessment (DA) scores, provided mostly by non-expert raters, on a scale from 0 to 100. WMT data with DA scores is available for years between 2015 and 2023, and covers nearly 50 language pairs.

2. MQM scores (Lommel et al., 2014; Freitag et al., 2021) on a scale from 0 to 25 (or uncapped in the most recent years), where lower is better. The scores are derived from professional translator annotations of error spans, including error severities and categories, where, generally, each minor error and each major error contribute 1 and 5 to the score, respectively. MQM annotations are only available for a limited number of language pairs (en-de, en-es, en-ru, en-zh, he-en, zh-en, ja-zh) for WMT data starting from 2020.

3. ESA scores (Kocmi et al., 2024b), which combine the first two approaches in that the expert raters annotate error spans and severities, yet they provide an overall quality score on a scale from 0 to 100 as well. These were only introduced in WMT24, so there is only one year worth of ESA data.

We use both DA and MQM scores for training MetricX, our score prediction system, and MQM error span annotations for training GemSpanEval. In general, we reserved data from WMT24 – both MQM (Freitag et al., 2024) and ESA (Kocmi et al., 2024a) – for validation of our models, with the exception of one submission to the error span detection task. We provide more details on the training and validation sets in §3 for MetricX and in §4 for GemSpanEval.

3 Quality Score Prediction: *MetricX-25*

Our MetricX-25 submissions to the Quality Score Prediction subtask are based on the successful MetricX-24 (Juraska et al., 2024; Freitag et al., 2024), with several modifications and improvements. The biggest one among them is the switch from mT5 (Xue et al., 2021) to Gemma 3 as the backbone model. We start this section by providing an overview of the similarities and differences between MetricX-24 and MetricX-25, then give more details on the training and evaluation data, and finally describe our experiments and the MetricX-25 systems we submitted to the shared task.

3.1 MetricX Overview

MetricX is a regression model trained to predict a quality score for a machine translation given the source segment and/or a reference translation. As of the '24 version, there are no separate models for reference-based and reference-free prediction; instead, a single model is trained on a mixture of examples: (1) with both the source and a reference, (2) with the reference omitted, and (3) with the source omitted.

The model is trained on translation evaluation data in two stages. It is first fine-tuned on z-normalized DA scores, then further fine-tuned on a mixture of MQM scores and raw DA scores. In both stages, a small proportion of synthetic training data, generated from WMT data, is included to help MetricX models recognize certain types of bad translations that are insufficiently represented in the standard WMT data. These synthetic examples cover cases such as over- and undertranslation, fluent but unrelated translation, and missing punctuation. We refer the reader to Juraska et al. (2024) for the full list of synthetic example categories and details on how the data was constructed.

3.2 What Is New in MetricX-25?

Initialization model. The model we initialize MetricX-25 from is Gemma 3 12B, a state-of-the-art multilingual open-weights model similar in size to the 13B-parameter mT5-XXL used in all previous versions of MetricX. In contrast to mT5-based MetricX models, MetricX-25 uses an encoder-only architecture with a regression head on top. Specifically, MetricX-25 is a fine-tuned Gemma Encoder (Suganthan et al., 2025) with mean pooling and no uptraining, and with the encoder's weights initialized from the corresponding

Czech source:
```Připadalo mi, že na mě dýchnul závan z hrobu.```

English (United Kingdom) reference:
```It was like having felt a draught from a grave.```

English (United Kingdom) translation:
```It was like having felt a draft from a grave.```

Score type: MQM

Figure 1: Example MetricX-25 model input.

decoder weights of Gemma 3. Besides the major differences in architectures and pretraining corpora and strategies, Gemma 3 also has the advantage of supporting significantly longer context windows (up to 128K tokens) than mT5 and has an even wider language support (over 140 languages, compared to mT5's 101).

Language indication. For MetricX-25, we augment the model input with source and target language information. The motivation behind this is to help MetricX recognize when an untranslated source (or portion of it) is appropriate, as well as help it handle quality assessment of translations from one language dialect to another without incorrectly assuming the translation is largely untranslated. Thus, we include the country information too if the locale is indicated in the data (e.g., ar_EG) and the language has multiple major dialects, such as Arabic (Egyptian, Modern Standard Arabic, etc.) or Portuguese (Brazilian, European, etc.). Figure 1 shows a full example of a model input.

Input format. Given the dual nature of human evaluation in the WMT25 shared task – MQM for some language pairs and ESA for others – we also add a score type indication to the model input, so that it learns to predict both types of quality score. We use the "MQM" score type for the MQM training and evaluation data, and "ESA" for the DA training and ESA evaluation data. Considering the possibly multi-paragraph segments in the official test sets, we also enclose each segment between triple backticks and separate input segments with double newline characters (see Figure 1).[1]

2-way hybrid input mode. We changed the training recipe for MetricX-25 by only including reference-only examples in the first stage of fine-tuning. In the second stage, we fine-tune the model on two types of examples only: (1) source-only, and (2) with source and reference both. The reason is twofold. First, MQM scores were produced in a source-based fashion, without a reference being available at all, so training examples with an MQM score but no source segment might not provide an accurate signal to the model. We verified experimentally that omitting these examples in the second stage does not have a significant negative impact on reference-only prediction performance. Second, in all the evaluation scenarios of the quality score prediction task, source segments are available, so there is little reason to distract the model with source-free training examples in the second stage of fine-tuning.

Score clipping. Earlier versions of MetricX had MQM scores in the training data, as well as the output scores, clipped to the [0, 25] range. However, with the switch to document-level segments over the past two years of the shared task, and the fact that MQM scores in human evaluation are therefore no longer capped at 25, we also drop the score clipping in MetricX-25, allowing for output scores greater than 25, which we expect to improve the correlation with MQM scores for long segments.[2]

3.3 Experimental Setup

Training data. In the first stage of fine-tuning MetricX-25, we use z-normalized DA scores from WMT15–23, with the into-English subset of WMT21 omitted due to its low quality (Juraska et al., 2023). Furthermore, the DA z-scores are aggregated per segment, negated, and finally clipped to the $[-1.0, 1.0]$ range, as shown by Juraska et al. (2023) to yield the best performance. We also incorporate a small proportion of the synthetic training data introduced in Juraska et al. (2024). In this stage, we do not include any score type indication in the input. In the second fine-tuning stage, we mix an equal proportion of the same DA data as above (with "ESA" indicated as the score type) and MQM data from WMT20–23 (with "MQM" score type), along with synthetic data included in each group of examples. In contrast to the first stage, however, we use raw DA scores rescaled to the MQM scale here, so the model does not have to learn two different scales, only distributions. We

[1] We experimented with including a preamble with instructions on the quality assessment task in the input too, but it did not improve the performance, perhaps except for the first few steps of fine-tuning.

[2] This causes a small discrepancy with the synthetic training data, which uses a fixed range of [0, 25], with many of the examples having a score of 25 assigned to them. Nevertheless, most of the synthetic examples are sentence-level, so an MQM score of 25 is reasonable for very bad translations.
```

then rescale the output scores to their respective ESA and negative MQM scales, as expected for the evaluation on the official test set, in postprocessing.

Meta-evaluation. Our validation set, which we use to pick the best model checkpoints, consists of both the MQM and ESA data from WMT24. To evaluate our models' performance, we calculate to what degree their predicted scores agree with the human judgments of translation quality. For segment-level correlation we use the tie-calibrated pairwise accuracy introduced by Deutsch et al. (2023), while at the system level we calculate soft pairwise accuracy (SPA; Thompson et al. 2024). These were the two primary meta-evaluation metrics used in the WMT24 Metrics Shared Task (Freitag et al., 2024). We use the same checkpoint selection strategy as for MetricX-24, averaging over all three MQM language pairs of WMT24, and downweighting the system-level component due to its larger variance.

Implementation details. MetricX-25 is implemented in TensorFlow (Abadi et al., 2015), and all of our submitted MetricX-25 systems are based on the Gemma 3 variant with 12B parameters. We defer further implementation details to Appendix A.

3.4 Results and Submission Details

Here we present the results of our experiments, focusing on assessing the impact of the combined MQM/ESA score prediction and comparing the performance of the Gemma-based MetricX-25 with that of the similarly-sized mT5-based MetricX-24. Due to limited resource availability, we were only able to run each experiment with one random seed.

3.4.1 Combining MQM/ESA Score Prediction

We start by examining the effects of mixing DA and MQM training data, together with using the score type indicators in the input. As a reminder, raw DA scores are used to train the model to predict ESA scores, since they both use the same scale and follow similar distributions. The first 3 rows of Table 1 compare the combined fine-tuning with fine-tuning on DA data only and MQM data only. We can see that the "DA + MQM" model (row 3) performs on par with the "MQM only" model (row 2) on the MQM validation sets and on par with the "DA only" model (row 1) on the ESA validations sets. This demonstrates that the model can effectively learn to predict both types of scores without sacrificing performance in either of them.

Our next observation is that simple two-stage fine-tuning (DA first then MQM) also achieves this goal, except for the system-level performance on ESA, which is around 2 points behind both the "DA + MQM" and the "DA only" model (compare row 4 against rows 3 and 1). Finally, we show that by combining two-stage fine-tuning and DA/MQM data mixing (row 5), we significantly boost the system-level performance on the ESA sets, while maintaining or further improving the performance on the MQM sets, as well as maintaining the segment-level performance.

3.4.2 MetricX-25 Submissions

Table 2 summarizes our MetricX-25 submissions to the quality score prediction task and compares them against MetricX-24 – one of the three top-performing systems in last year's shared task – as a baseline. The submissions only differ in the combination of examples they were trained on (with or without source/reference) and thus their expected input: the primary submission is a hybrid system, whereas the two secondary submissions are a purely quality estimation (QE) and a purely reference-based system, respectively.

As the table shows, all the MetricX-25 submissions (rows 3–6) significantly outperform the MetricX-24 baseline (rows 1–2) at the segment level, but the results are mixed at the system level. The ja-zh language pair exhibits by far the largest improvement, suggesting that Gemma 3 has a stronger understanding of Japanese and/or Chinese than mT5. Our expectation is that this is true for many more languages, making MetricX-25 a more robust automatic evaluation metric than its mT5-based predecessors.

We chose the hybrid model as our primary submission, despite being slightly outperformed by the reference-based variant (compare rows 6 and 4), because of its input flexibility. Since the official test set consists of language pairs both with and without references available, the reference-based model would likely perform poorly on the latter. Moreover, the majority of the challenge sets also do not provide references, so MetricX-25-Ref would not be able to participate in them.

4 Error Span Detection: *GemSpanEval*

In this section, we describe our submission to subtask 2: Error Span Detection. We denote our system *GemSpanEval*, as it is a span-level prediction model based on Gemma 3.

Training protocol	Segment-level pairwise accuracy				System-level soft pairwise accuracy			
	en-de	en-es	ja-zh	Avg(ESA)	en-de	en-es	ja-zh	Avg(ESA)
DA only	50.41	68.51	54.48	54.72	85.62	**82.13**	92.62	87.65
MQM only	54.71	68.80	56.36	54.20	85.55	78.56	89.94	86.45
DA + MQM	54.71	68.92	56.16	54.91	85.21	78.89	90.67	**87.80**
DA→MQM	55.40	69.11	57.90	55.00	85.91	78.12	**93.64**	85.74
DA→(DA + MQM)	**55.66**	**69.25**	**58.24**	**55.14**	**86.60**	77.92	92.88	87.08

Table 1: Meta-evaluation scores of reference-based models (non-hybrid) on the WMT24 validation set, using a variety of one- and two-stage fine-tuning protocols, with the → symbol indicating two stages. "DA + MQM" denotes the combination of DA and MQM scores, with a score type indication provided in the input. The correlation scores are shown for all 3 MQM language pairs individually, and for the 9 ESA language pairs averaged.

MetricX variant	Segment-level pairwise accuracy				System-level soft pairwise accuracy			
	en-de	en-es	ja-zh	Avg(ESA)	en-de	en-es	ja-zh	Avg(ESA)
24-Hybrid-QE	52.60	68.50	53.00	–	**87.80**	78.90	87.50	–
24-Hybrid	53.20	68.50	53.90	–	87.40	**79.90**	89.70	–
25-QE†	54.97	**69.42**	57.21	54.34	85.45	78.29	91.34	84.91
25-Ref†	**55.66**	69.25	**58.24**	**55.14**	86.60	77.92	**92.88**	87.08
25-Hybrid-QE	54.83	69.31	56.66	54.11	85.42	77.74	91.59	86.32
25-Hybrid*	55.45	69.14	57.72	54.87	85.82	77.00	92.00	**87.61**

Table 2: Performance of our MetricX-25 submissions compared to the MetricX-24 baseline on WMT24. "Hybrid" and "Hybrid-QE" rows correspond to the same model, only evaluated with and without references, respectively. The correlation scores are shown for all 3 MQM language pairs individually, and for the 9 ESA language pairs averaged. *Primary submission. †Secondary submissions.

4.1 Overview

Last year's shared task on error span detection (Zerva et al., 2024), despite low participation, showed that span-level error prediction remains a challenging task. Specifically, Shan et al. (2024) found that generative methods based on LLMs, such as GPT-4o-mini (Hurst et al., 2024) or Tower-Instruct-7B (Alves et al., 2024), still lag behind encoder-only models like CometKiwi (Rei et al., 2022b). Our shared task submission aims to explore how far we can push generative models at the error span detection task when using a recent strong multilingual open-weights model, in our case Gemma 3 27B, fine-tuned for the error span detection task.

4.2 Error Span Detection

We adapt the AutoMQM setup (Fernandes et al., 2023) by predicting MQM error spans in JSON format similar to Finkelstein et al. (2024). LLMs tend to be good at producing valid JSON output, making parsing the structured output straightforward. Each error contains the text span of the machine translation or the source segment (for omissions) along with a severity and a category. Note that for the error span detection task the category and source errors, such as omissions, are not used.

While models are able to identify and extract spans, we also need to find the spans in the original text for this task. A simple way to do so is to perform a string search. This is successful for most error spans, as they are unique substrings of the source or the machine translation. However, there is also a considerable portion of error spans that are not unique, often related to punctuation or short frequent words. For example, in the WMT24 en-de MQM data, 21% of error spans are not unique. Note though, that the problem is less pronounced for the shared task evaluation, as it is based on the character F1 score, to which short spans contribute less than long spans.

To be able to uniquely identify short spans, we modify the model response to include additional context for any span that is not unique. We expand the context to the previous and next word by looking for the previous and next space character. For Chinese and Japanese, we extend the context by one character at a time. The context is extended until we find a unique substring. See Figure 2 for an example of translation error spans with context in JSON format, and Figure 3 in the Appendix for the full prompt and output. The rest of the prompt and format follows Finkelstein et al. (2024). We do not use ICL examples, as we train a dedicated model. All data is presented twice: once with the

```
English source:
‘‘‘I have not made use of the timer,
    preferring to turn them on and off
    myself. I can see this feature as
    useful in an office setting with
    houseplants or if on vacation‘‘‘
German machine translation:
‘‘‘Ich benutze den Timer nicht, sondern
    schalte ihn lieber selbst ein und
    aus. Ich sehe diese Funktion als
    nützlich im Büro mit Zimmerpflanzen
    oder im Urlaub.‘‘‘
```

Response:

```
[
  {"span": "im", "span_with_context":
      "nützlich im Büro", ...},
  {"span": "ihn", ...},
  {"span": "mit", ...}
]
```

Figure 2: Example translation with non-unique error spans, where span context text is included.

reference translation and once without. This allows us to evaluate the model as a QE and as a reference-based model both.

4.3 Experimental Setup

For development, we train on MQM data from WMT20–23 and evaluate on WMT24 MQM data. We evaluate using character-level F1 score, which was used as the shared task metric in previous years (Zerva et al., 2024) and also the current year. The metric takes into account error severities and gives partial credit for predicting an error span with the wrong error severity. For the submission system, we include the en-de and ja-zh WMT24 MQM data in training as well, holding out en-es for evaluation. We include the latest data in order to fine-tune the model on longer segments too. Before WMT24, only WMT23 en-de provided paragraph-level data, while all other data is at the sentence level. Additionally, we hope to increase the coverage of translation errors of modern LLM-based translation models.

We fine-tune a 27B Gemma 3 model using Kauldron SFT tooling.[3] We use the Adafactor (Shazeer and Stern, 2018) optimizer with a learning rate of 0.0001 and a batch size of 64, running for 20K steps, which covers a little under 2 epochs of the training data. As baselines we report XCOMET-XXL (Guerreiro et al., 2024) scores. XCOMET is a strong encoder-only model that was trained to rate segments and also label translation error token spans. Additionally, we evaluate Gemma 3 27B without fine-tuning, prompting it to produce JSON error spans, as shown in Figure 3. Note that this baseline does not include the span context for short spans.

[3] https://kauldron.readthedocs.io/en/latest/

4.4 Results and Submission Details

The experimental results for the span-level error prediction task are shown in Table 3. GemSpanEval-QE v1 denotes an initial training run for just 10K steps and without span context for short spans. GemSpanEval-QE and GemSpanEval are the QE and reference-based evaluations, respectively, of a model trained without WMT24 data. The final row adds WMT24 en-de and ja-zh data. Therefore, for the last row, we should only take the results as an indication of how much of WMT24 has been memorized, not of how good the model is, except for en-es. The final shared task submission is based on the model that was trained *with* WMT24 data. The primary submission uses the reference while the secondary submission is reference-free. Both submissions use the same model and just differ in whether references are shown at inference time. Note that, when preparing the submission data, we ran into model repetition problems that resulted in invalid JSON output. For these cases we fell back to the original model, GemSpanEval-QE v1. For the shared task submission we find 22% of errors spans to not be unique, while this corresponds to only 5% of the characters of error spans, as non-unique spans tend to be short. Consequently, for WMT24 en-de this leads to a marginal F1 improvement of 0.08.

System	en-de	en-es	ja-zh	Avg.
XCOMET-XXL-QE	24.28	10.11	14.30	16.23
XCOMET-XXL	**25.43**	11.02	24.94	**20.46**
Gemma 3 27B	17.94	8.19	**28.42**	18.18
GemSpanEval-QE v1	17.51	14.43	22.75	18.23
GemSpanEval-QE	20.85	13.06	24.72	19.54
GemSpanEval	21.79	13.73	25.28	20.27
+ WMT24 train	27.26	**14.37**	37.09	26.24

Table 3: WMT24 character level F1 scores for the error span prediction task. Numbers where we train on the development set are grayed out but kept for reference.

From the experimental results during development in Table 3, we see that the encoder-only XCOMET-XXL model is a strong baseline showing the best result for en-de. The Gemma 3 baseline that was not fine-tuned also shows generally good performance, even achieving the best

result of all evaluated systems for ja-zh, better than the fine-tuned model. This is likely due to the task being difficult and highly dependent on rater behavior, which varies significantly across years and languages. Using references consistently achieves higher character F1 across all three language pairs. Surprisingly, the model trained with WMT24, while showing the best results in terms of character F1, still shows a significant gap, despite seeing the test data during training for en-de and ja-zh. This shows that ∼2 epochs in the current training setup was not enough to completely memorize the training data. For the held-out language, en-es, we see the best score across all settings, leading to the decision to use the model trained with WMT24 as our submission to the shared task.

5 Related Work

For decades, right until the recent advent of LLMs, the most widely adopted automatic evaluation metrics would express the predicted machine translation quality as a scalar score. This is the case for metrics ranging from simple lexical overlap metrics, such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), to learned metrics including BLEURT (Sellam et al., 2020; Pu et al., 2021), COMET (Rei et al., 2020, 2022a) and MetricX (Juraska et al., 2023, 2024). The feasibility of prompting general-purpose LLMs to score translations has also been studied (Kocmi and Federmann, 2023b; Leiter et al., 2023; Leiter and Eger, 2024) and this approach has been shown to be competitive with fine-tuned dedicated models, especially in system-level performance. Among recent methods, there has been an increasing proportion of metrics providing structured (Perrella et al., 2022; Kocmi and Federmann, 2023a; Fernandes et al., 2023; Guerreiro et al., 2024) or natural language explanations (Xu et al., 2023) for the predicted scores, most of which are based on LLMs.

Since the era of lexical metrics, which require one or more reference translations to evaluate a machine translation, metrics have evolved to rely increasingly more on the source segment (Rei et al., 2020, 2022a; Juraska et al., 2024). This is enabled by the multilingual pretrained models they are typically built on top, such as XLM-R (Conneau et al., 2020) or mT5 (Xue et al., 2021). In fact, *reference-free* (or *quality estimation*; QE) metrics do not lag far behind their *reference-based* counterparts anymore, as evidenced by the most recent WMT Metrics shared tasks (Freitag et al., 2023, 2024). That being said, high-quality references do provide added value to automatic metrics in most cases, helping them make even more accurate quality predictions. Therefore, metrics nowadays typically employ a unified (or hybrid) input approach, allowing them to make a reference-based prediction whenever a reference is available, and a QE prediction otherwise (Wan et al., 2022; Guerreiro et al., 2024; Juraska et al., 2024). This is the case with our primary MetricX-25 and GemSpanEval submissions as well.

Current approaches to error span annotation are often based on encoder-only models predicting error severities per token. One such approach is COMETKIWI (Rei et al., 2022b, 2023) or, more recently, XCOMET (Guerreiro et al., 2024). In last year's Error Span Detection shared task, COMETKIWI was shown to be a competitive baseline, coming in ahead of the (single) submitted system (Zerva et al., 2024). Kocmi and Federmann (2023a) showed that GPT-4 can simply be prompted to produce MQM error spans, denoting the method GEMBA-MQM. Fernandes et al. (2023) showed that fine-tuning for the generative error span prediction can improve performance compared to prompting LLMs alone. Recent translation-oriented LLMs, such as Tower (Alves et al., 2024; Rei et al., 2025) also include generative error span prediction as part of the supported tasks in the training data.

6 Conclusion

We introduced MetricX-25 and GemSpanEval, our submissions to the WMT25 Evaluation Shared Task, both built upon the Gemma 3 foundation model, but in very different ways. We demonstrated that MetricX-25, adapting Gemma 3 to an encoder-only architecture, can be trained to effectively predict ESA and MQM quality scores and significantly outperforms its predecessor, MetricX-24, in segment-level performance. For error span detection, our generative model, GemSpanEval, proved to be competitive with a strong sequence-tagging baseline. Additionally, we showed how error span context can be used to identify unique error spans. Our work demonstrates that a strong multilingual foundation model, such as Gemma 3, can successfully be used for both regression-based and generative translation evaluation metrics.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Mara Finkelstein, Dan Deutsch, Parker Riley, Juraj Juraska, Geza Kovacs, and Markus Freitag. 2024. From jack of all trades to master of one: Specializing llm-based autoraters to a test set. *arXiv preprint arXiv:2411.15387*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil

Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Christoph Leiter and Steffen Eger. 2024. PrExMe! large scale prompt exploration of open source LLMs for machine translation and summarization evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506, Miami, Florida, USA. Association for Computational Linguistics.

Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. The Eval4NLP 2023 shared task on prompting large language models as explainable metrics. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the*

Tenth Workshop on Statistical Machine Translation, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Weiqiao Shan, Ming Zhu, Yuang Li, Mengyao Piao, Xiaofeng Zhao, Chang Su, Min Zhang, Hao Yang, and Yanfei Jiang. 2024. HW-TSC 2024 submission for the quality estimation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 535–540, Miami, Florida, USA. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Paul Suganthan, Fedor Moiseev, Le Yan, Junru Wu, Jianmo Ni, Jay Han, Imed Zitouni, Enrique Alfonseca, Xuanhui Wang, and Zhe Dong. 2025. Adapting decoder-based language models for diverse encoder downstream tasks. *arXiv preprint arXiv:2503.02656*.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

A MetricX-25 Implementation Details

With the input context length of 4096 SPM tokens, each training run requires 64 TPUs. Using a batch size of 128, we train our models for 10K steps in the first stage, using a learning rate of 0.00005 with a cosine decay following 100 steps of linear warm-up. We then fine-tune the best checkpoint for another 10K steps in the second stage, lowering the peak learning rate to 0.00001.

Prompt

```
You are an annotator for the quality of machine translation. Your task is to
    identify errors and assess the quality of the translation.
Based on the source segment, human-generated reference translation, and machine
    translation surrounded with triple backticks, identify error types in the
    translation and classify them. The categories of errors are: accuracy
    (addition, mistranslation, omission, untranslated text), fluency (character
    encoding, grammar, inconsistency, punctuation, register, spelling), style
    (awkward), terminology (inappropriate for context, inconsistent use),
    non-translation, other, or no-error.
Each error is classified as one of three severities: critical, major, and minor.
    Critical errors inhibit comprehension of the text. Major errors disrupt the
    flow, but what the text is trying to say is still understandable. Minor errors
    are technically errors, but do not disrupt the flow or hinder comprehension.

Make sure your response is a strict and valid json object that could be parsed with
    json.loads() in python.

English source:
‘‘‘The lights are dimmable, but I use the strongest setting only. I have not made
    use of the timer, preferring to turn them on and off myself. I can see this
    feature as useful in an office setting with houseplants or if on vacation‘‘‘
German machine translation:
‘‘‘Die Lichter sind dimmbar, aber ich benutze nur die stärkste Einstellung. Ich
    benutze den Timer nicht, sondern schalte ihn lieber selbst ein und aus. Ich sehe
    diese Funktion als nützlich im Büro mit Zimmerpflanzen oder im Urlaub.‘‘‘
```

Response:

```
[
  {"span": "im", "severity": "minor", "category": "accuracy/mistranslation",
     "span_with_context": "nützlich im Büro"},
  {"span": "ihn", "severity": "minor", "category": "accuracy/mistranslation"},
  {"span": "mit", "severity": "minor", "category": "accuracy/mistranslation"}
]
```

Figure 3: Example prompt and response for AutoMQM error span identification. We omit the error span attribute `is_source_error` for brevity. Each span that is not unique receives an additional attribute `span_with_context`.

HW-TSC's submissions to the WMT 2025 Segment-level quality score prediction Task

Yuanchang Luo*, Jiaxin GUO*, Daimeng Wei*, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Xiaoyu Chen and Hao Yang
Huawei Translation Services Center, Beijing, China
{luoyuanchang1,guojiaxin1,weidaimeng}@huawei.com

Abstract

This paper presents the submissions of Huawei Translate Services Center (HW-TSC) to the WMT 2025 Segment-level quality score prediction Task. We participate in 16 language pairs. For the prediction of translation quality scores for long multi-sentence text units, we propose an automatic evaluation framework based on alignment algorithms. Our approach integrates sentence segmentation tools and dynamic programming to construct sentence-level alignments between source and translated texts, then adapts sentence-level evaluation models to document-level assessment via sliding-window aggregation. Our submissions achieved competitive results in the final evaluations of all language pairs we participated in.

1 Introduction

Recent advances in large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Yang et al., 2024) have opened new possibilities for document-level machine translation (doc-mt) (Kim et al., 2019; Maruf et al., 2022; Fernandes et al., 2021). Leveraging their robust language generation capabilities and profound contextual understanding, LLMs can produce translations that are more natural, fluent, and semantically coherent. These models have demonstrated remarkable proficiency in processing long-form texts, thereby significantly enhancing the quality of document-level translation.

However, this approach also introduces several challenges. Since LLMs translate entire documents holistically rather than processing sentences sequentially, the output may suffer from issues such as over-translation (excessive paraphrasing) or under-translation (omissions). Furthermore, the absence of sentence-level alignment between source and target texts—combined with the inherent length of both—makes it difficult to assess translation quality accurately. Robust evaluation methods for document-level machine translation (MT) remain an unresolved critical problem.

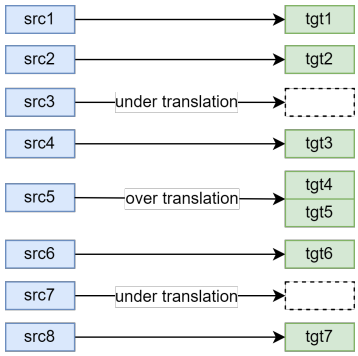


Figure 1: src_3 and src_7 lack corresponding translations in T, while src_5 aligns with a combined $tgt_4 + tgt_5$ segment.

While human evaluation remains the gold standard for assessing translation quality due to its nuanced understanding of language and context, it faces inherent limitations in scalability, subjectivity, and cost-efficiency, particularly for large-scale document-level translation tasks. Automated metrics like BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020a,b), though capable of capturing semantic nuances and demonstrating strong correlation with human judgments, are constrained by input length restrictions and their reliance on sentence-level alignment between source and reference texts. While (Vernikos et al., 2022) pioneered the adaptation of these metrics to document-level translation evaluation, its applicability remains severely constrained by its fundamental requirement for perfect sentence-level alignment among source texts, translations, and reference translations. This strict one-to-one correspondence prerequisite significantly limits its practical utility in real-world scenarios where such ideal alignments rarely exist. Recent attempts to leverage large language models (LLMs) as eval-

*These authors contributed equally to this work.

Proceedings of the Tenth Conference on Machine Translation, pages 969–973
November 8-9, 2025

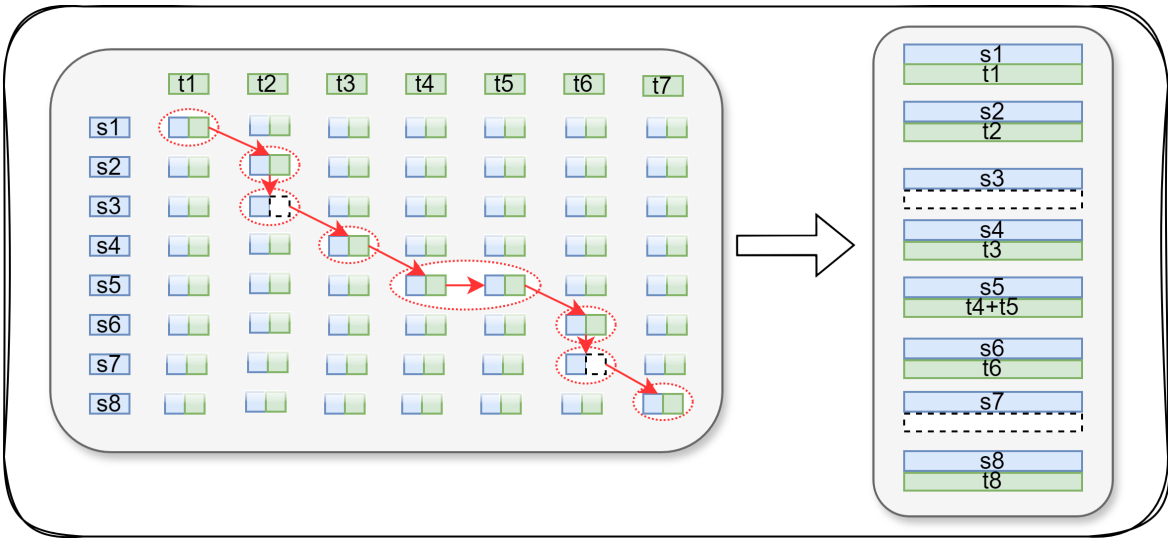


Figure 2: For the segmented text pair (8 source fragments and 7 target fragments), we first compute a full 8×7 score matrix using COMET KIWI to evaluate all possible pairwise alignments(subfigure a). We then apply dynamic programming to identify the optimal alignment path (visualized as the red trajectory in Figure). This optimization yields final sentence-level alignments, resulting in 8 properly aligned source-target pairs as demonstrated in subfigure (b).

uators through carefully designed prompts show promising alignment with professional human assessments across multiple dimensions including accuracy, fluency, and stylistic consistency (Gu et al., 2025). However, these methods suffer from high computational costs, sensitivity to training data biases, and instability across different prompts or model runs, raising concerns about their reliability and reproducibility for practical applications.

In this work, we employ an innovative alignment algorithm to automatically construct sentence-level alignment between source and translated texts. Our approach(Guo et al., 2025) involves: (1) sentence segmentation of source and target texts, (2) alignment metric computation, (3) anchoring of source text segmentation information, and (4) reconstructed target text segmentation (including merging and gap filling). By subsequently applying sliding-window-based sentence-level evaluation, we achieve document-level assessment effectiveness, thereby successfully adapting sentence-level pretrained model evaluation methods to document translation.

2 Approach

2.1 Alignment

Since our source text, translation, and reference translation are all document data, the sentence-level alignment between the source text and translation that we automatically construct can be divided into the following three parts:

- Sentence segmentation: Segment both original and translated texts into sentence sequences.
- Calculate alignment metrics: Measure alignment similarity between original and translated sentences using metrics like COMET KIWI (Rei et al., 2022) or LABSE (Feng et al., 2022).
- Reconstruct translated text segmentation: Based on the original text's segmentation, reconstruct the translated text's segmentation, involving possible merging or filling gaps. This is done using a dynamic programming algorithm.

As shown in Figure 2, for a source text S and its target translation T, we first perform sentence segmentation using spaCy [1], yielding m source sentences $S = (s_1, s_2, ..., s_m)$ and n target sentences $T = (t_1, t_2, ..., t_n)$. For these $m \times n$ sentence pairs, we compute a KIWI matrix $KIWI_{m \times n}$ using COMET KIWI. When $m = n$ with one-to-one correspondence, the diagonal path of this matrix should yield the maximum values. In document-level translation scenarios, the number of source segments and target segments typically differs

[1] https://spacy.io/

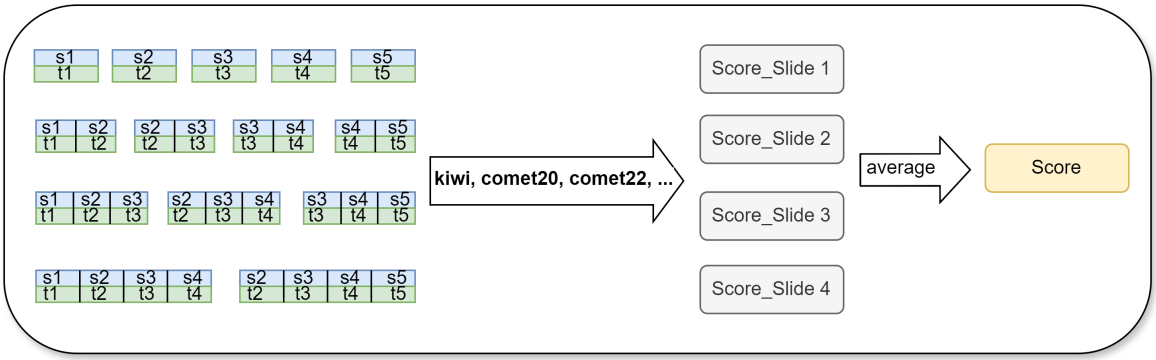


Figure 3: For the reconstructed source-target pairs, Compute Score Slide 1 on 5 original aligned pairs. Generate 4 concatenated pairs using window size 2 to calculate Score Slide 2 Generate 3 concatenated pairs using window size 3 to calculate Score Slide 3 Generate 2 concatenated pairs using window size 4 to calculate Score Slide 4. The final document-level metric is derived by averaging these four window-level scores, providing comprehensive coverage of local and contextual translation quality.

$(m \neq n)$. Nevertheless, we can identify an optimal alignment mapping $T = (t_1, t_2, ..., t_m) = F(s_1, s_2, ..., s_n)$ - represented as the optimal path in our framework - that maximizes the COMET KIWI score.

This alignment task can be abstracted as a path optimization problem: Given an $[mn]$ matrix where each cell (i, j) contains a score value, we seek the optimal path from $(0, 0)$ to $(m-1, n-1)$ under the following constraints:

- **Monotonicity Constraint**: y-coordinate must increase by exactly 1 at each step $(\forall t, y_{t+1} = y_t + 1)$. x-coordinate must increase by a non-negative integer $(\forall t, x_{t+1} \geq x_t)$
- **Boundary Conditions**: Path originates at the top-left corner $(0, 0)$ and terminates at the bottom-right corner $(m-1, n-1)$
- **Optimization Objective**: Maximize the cumulative score:

$$argmax_p \sum_{(x,y)\in p} matrix[x][y] \quad (1)$$

Using the dynamic programming algorithm, we can obtain a translation whose segmentation aligns one-to-one with the source text, as well as the segmentation information of the reference translation.

2.2 Sliding Evaluation

After obtaining the alignment information in the previous step, we follow a procedure similar to (Raunak et al., 2024), calculating sentence-level scores using a sliding window approach. As illustrated in Figure 3, for m source sentences $S = (s_1, s_2, ..., s_m)$ and their aligned translations $T' = (t'_1, t'_2, ..., t'_m)$, given a window size n, we compute m groups of sentence-level evaluation metrics, each incorporating $n-1$ preceding sentences as contextual information. The mean of these scores serves as the document-level evaluation result, expressed formally as follows:

$$\frac{1}{n}\sum_{i=1}^{n} f_i(S, T') \quad (2)$$

Where f_i corresponds to the Slide Score measured when the window is i, corresponding to Score Slide i in Figure 3.

3 Results

We participated in all language pair competitions within the Segment-level Quality Score Prediction Task, which included a total of 16 language pairs. After the alignment phase, we obtained new sentence-level text pairs corresponding to each paragraph text pair. At this point, we conducted respective predictions using wmt22-cometkiwi-da(**ASD-KIWI**), wmt23-cometkiwi-da-xl(**ASD-KIWI-XL**), and wmt23-cometkiwi-da-xxl(**ASD-KIWI-XXL**), with the results shown in Table 1. As shown in the Table 1, **ASD-KIWI-XL** demonstrates superior correlation to **ASD-KIWI** across most of the 16 language pairs, indicating that post-alignment sentence-pair quality scoring plays a critical role. While larger parameter models generally achieve better performance (as evidenced by KIWI-XL's gains), this trend is not absolute—**ASD-KIWI-XXL** fails to further outperform **ASD-KIWI-XL**.

Languages Pairs	ASD-KIWI	ASD-KIWI-XL	ASD-KIWI-XXL	ASD-KIWI-ENSEMBLE
EN-ZH	0.6800	0.6467	0.5600	**0.7467**
CS-UK	0.7382	0.5418	0.7164	**0.7818**
EN-KO	0.7067	0.7600	0.7267	**0.7733**
EN-IT	0.7169	0.600	0.5077	**0.7169**
EN-ET	0.7018	0.7164	0.6436	**0.7455**
EN-BHO	0.9316	0.9031	0.8348	**0.9316**
EN-IS	**0.8667**	0.7667	0.7400	0.7933
EN-SR	0.8974	0.8575	0.8519	**0.9031**
CS-DE	0.719	0.5820	0.6676	**0.7418**
EN-RU	0.6710	0.6017	0.3853	**0.8355**
EN-JA	**0.7933**	0.6533	0.4933	0.7667
EN-AR	0.8551	0.8696	0.7681	**0.8551**
EN-UK	0.7524	0.7238	0.5048	**0.8190**
EN-MAS	**0.7628**	0.5652	0.5889	0.5968
EN-CS	0.5942	0.6087	0.5362	**0.6957**
JA-ZH	**0.7245**	0.5042	0.6177	0.6978

Table 1: Results for 16 Languages Pairs in the Segment-Level Quality Score Prediction Task

To leverage both models, we propose an ensemble method that averages the per-sentence scores of **ASD-KIWI** and **ASD-KIWI-XL**. Empirical results confirm that **ASD-KIWI-ENSEMBLE** achieves the best overall performance.

4 Conclusion

This paper presents the methodology behind HW-TSC's submission to the WMT 2025 Segment-Level Quality Score Prediction Task. Our approach integrates sentence segmentation tools and dynamic programming algorithms to construct sentence-level alignments between source and translated texts, then adapts sentence-level evaluation models to document-level assessment through sliding-window aggregation. By incorporating an ensemble strategy, our method achieved the highest correlation scores across all 16 languages in this task.

References

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6467–6478. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Xiaoyu Chen, Zhanglin Wu, Huan Yang, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, and 1 others. 2025. Align-then-slide: A complete evaluation framework for ultra-long document-level machine translation. *arXiv preprint arXiv:2509.03809*.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 24–34. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: reference-free evaluation for machine trans-

lation using a sliding document window. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 205–211. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 911–920. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 118–128. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

UvA-MT at WMT25 Evaluation Task: LLM Uncertainty as a Proxy for Translation Quality

Di Wu **Christof Monz**
Language Technology Lab
University of Amsterdam
{d.wu, c.monz}@uva.nl

Abstract

This year, we focus exclusively on using the uncertainty quantification as a proxy for translation quality. While this has traditionally been regarded as a form of **unsupervised** quality estimation, such signals have been overlooked in the design of the current metric models—we show their value in the context of LLMs. More specifically, in contrast to conventional unsupervised QE methods, we apply recent calibration technology (Wu et al., 2025b) to adjust translation likelihoods to better align with quality signals, and we use the **single** resulting model to participate in both the general translation and QE tracks at WMT25.

Our offline experiments show some advantages: 1) uncertainty signals extracted from LLMs, like Tower or Gemma-3, provide accurate quality predictions; and 2) calibration technology further improves this QE performance, sometimes even surpassing certain metric models that were trained with human annotations, such as CometKiwi. We therefore argue that uncertainty quantification (confidence), especially from LLMs, can serve as a strong and complementary signal for the metric design, particularly when human-annotated data are lacking. However, we also identify limitations, i.e., its tendency to assign disproportionately higher scores to hypotheses generated by the model itself.

1 Introduction

In this paper, we describe the details of our submission to the WMT 2025 MT evaluation subtask-1, i.e., segment-level Quality Estimation (QE), which includes 16 translation directions. This year, we focus exclusively on using the uncertainty quantification as a proxy for translation quality. While this has traditionally been regarded as a form of unsupervised quality estimation (Fomicheva et al., 2020), such signals have been overlooked in recent designs of metric models. In this competition, we aim to examine the strengths and weaknesses of this signal.

Previous unsupervised quality estimation focused on using the model's internal information to quantify the confidence/certainty of a given translation sentence pair, e.g., using likelihood, entropy, or uncertainty signals under a Monte Carlo (MC) dropout framework (Fomicheva et al., 2020). Notably, they are relying on signals derived from the model itself and are mostly training-free.

We apply recent calibration technology (Wu et al., 2025b) for this year's competition. Unlike traditional unsupervised QE, this method aims to calibrate translation likelihood with quality during training time.

By extensive experiments, several key advantages of calibrated models can be shown as follows:

- Translation quality can be substantially improved with limited training, e.g., 2K instances for each translation direction, and the effectiveness of maximum *a posterior* decoding, like beam search, can be better realized, showing strong promise for real-world use;
- **At the same time, it provides a unified view for optimizing translation quality and estimation, which matches our goal in this competition, i.e., using the model's confidence as a proxy for translation quality.**

Our offline experiments show that the resulting model's QE ability sometimes even surpasses some accurate metric models, like cometkiwi-22 (Rei et al., 2022)[1], without using any human-annotated data. We therefore argue that uncertainty quantification, especially from LLMs, can serve as a strong and complementary signal for the metric design, particularly when human-annotated data are lacking.

[1]https://huggingface.co/Unbabel/wmt22-cometkiwi-da

Proceedings of the Tenth Conference on Machine Translation, pages 974–983
November 8-9, 2025

However, our preliminary tests on this year's WMT25 test set also reveal limitations in using uncertainty signals for quality estimation—**notably, they tend to assign significantly higher scores to hypotheses generated by the model itself**—a pattern that is, to some extent, consistent with recent findings that LLM-as-a-judge systems tend to favor their own outputs (Panickssery et al., 2024).

In the next sections, we will briefly describe: 1) the framework of calibration, 2) offline experiments for both translation performance and our QE strategy by using the calibration approach, 3) the implementation for our submission at the WMT25-QE track, and 4) a discussion of the strengths and limitations of using model confidence for quality estimation.

2 Calibration Method

In this section, we briefly describe the calibration method to keep the paper self-contained; please refer to (Wu et al., 2025b) for details.

Formally, given a parameterized auto-regressive language model p_θ and a translation instruction x, the *log-likelihood* of a translation hypothesis y_i is denoted as $z_\theta(y_i|x) = \log p_\theta(y_i|x)$. Meanwhile, the quality of this translation can be defined as $q(y_i|x)$ where q represents any external quality evaluation model. When sampling hypotheses y from p_θ conditioned on x, both $z_\theta(y|x)$ and $q(y|x)$ can be viewed as random variables defined over the output space. This method is to calibrate the likelihoods of generated hypotheses with their quality to maximize the correlation between $z_\theta(y|x)$ and $q(y|x)$.

The calibration method uses the statistic, *Pearson correlation coefficient* $\rho(a, b)$, to quantify the correlation. Let $a, b : \mathcal{Y} \to \mathbb{R}$ be two real-valued functions defined over a domain $\mathcal{Y}$. The corresponding Pearson score between a and b is given by

$$\rho(a, b) = \frac{\mathbb{E}_{y\sim p}\left[(a(y) - \mu_a)\,(b(y) - \mu_b)\right]}{\sigma_a \sigma_b}, \quad (1)$$

where μ_a, μ_b and σ_a, σ_b denote expectations and standard deviations, respectively. This formulation computes the correlation by normalizing the expected product of their centered values. Due to its scale-invariance and ability to capture trend consistency, the Pearson correlation coefficient is widely used in translation metric meta-evaluation.

The calibration method calculates and optimizes ρ with respect to the likelihood of hypotheses $z_\theta(y|x)$ and the quality score $q(y|x)$. Practically, given the intractably large decoding space, it employs Monte Carlo sampling for approximation. For each source sentence x, we generate k hypotheses y_i ($i \in \{1, \ldots, k\}$) by repeatedly prompting a large language model θ with nucleus sampling, and compute the corresponding $z_\theta(y_i|x)$ and $q(y_i|x)$, and estimate the corresponding μ_z, μ_q and σ_z, σ_q. Accordingly, we define the Pearson-based loss using estimates under the nucleus-induced distribution $\tilde{p}$ as follows:

$$\mathcal{L}_{\mathrm{p}} = -\frac{1}{k} \sum_{\substack{i=1 \\ y_i\sim\tilde{p}_\theta(\cdot|x)}}^{k} \frac{z_\theta(y_i|x) - \mu_z}{\sigma_z} \cdot \frac{q(y_i|x) - \mu_q}{\sigma_q}. \quad (2)$$

It additionally introduces a supervised fine-tuning (SFT) term on the highest-scoring samples as a regularizer to ensure that the model's likelihood distribution remains grounded in high-quality translations, since the Pearson objective alone enforces correlation but does not constrain the absolute scale. The final loss for calibration is formulated as $\mathcal{L}_{\text{cal}} = \mathcal{L}_{\mathrm{p}} + \mathcal{L}_{\text{sft}}$.

An off-policy formulation can be obtained by trivially replacing the current model p_θ with an external model p_{θ^*} for sampling. Overall, by minimizing $\mathcal{L}_{\text{cal}}$, we encourage the Pearson score between z and q to increase. In practice, we use a gradient-based optimizer, Adam, to optimize θ for this goal, with gradients propagated through z_θ, μ_z, and σ_z. Despite its simplicity, several important characteristics are captured in this formulation:

- It models hypothesis qualities from a holistic view, enabling the model to make finer-grained distinctions in translation quality within the decoding space.
- It considers the value of translation quality by the metric function $q(\cdot|x)$, which is ignored in virtually all existing methods based on Bradley-Terry and Plackett-Luce, such as CPO.
- Pearson's correlation inherently applies normalization to a group of both likelihood and quality points. This normalization makes the objective invariant to scale and shift, thereby promoting stable and robust optimization across diverse input distributions.
- **The objective, i.e., the Pearson's score itself, is inherently shared with that of translation metric meta-evaluation, offering a unified perspective for both quality optimization and es-**

timation. Meanwhile, unlike other statistics like Spearman's or Kendall's scores, Pearson's coefficient is differentiable and thus suitable for gradient-based optimization frameworks.

3 Offline Evaluation

In this section, we demonstrate the effectiveness of the calibration method by applying it to the strong LLM-based translation system, i.e., Tower (Rei et al., 2024). We briefly show this method's capabilities in both (1) translation quality optimization and (2) quality estimation optimization.

Note that our online submission system is based on Gemma-3-12B (see Section 4), as it covers more languages for WMT25 than Tower.

3.1 Experimental Setups

Base model. We conduct experiments on Tower-7B, Tower-13B, and TowerMistral-7B models.

Calibration dataset. For the training set, we firstly merge all English sentences from the Flores-200 dataset (Costa-Jussà et al., 2022) in *dev* and *devtest* splits and use them as the source, consisting of 2,009 samples. We focus on off-policy experiments, where we use these sentences to construct translation prompts for each direction by calling an external strong translation model, rather than sampling from the base model itself. Here, we query `gpt-4o-mini`[2] 16 times per prompt, employing nucleus sampling with a temperature of 1.0 and a top-p of 0.98. The resulting bitexts are evaluated using CometKiwi-XXL to reflect corresponding quality scores. In this study, we only use this small-scale dataset to post-train our model. This is motivated by recent studies (Xu et al., 2023; Wu et al., 2024) showing that a few high-quality samples with a strong base model can significantly enhance the system's performance.

Training setups. For all experiments, we train models using LoRA (Hu et al., 2022) with rank 8, setting α to 32 and dropout to 0.05. Training uses a batch size of 32, gradient accumulation of 8 steps, and sequences capped at 512 tokens. To ensure robust results, we experiment with learning rates ranging from 1e-5 to 1e-4, reporting the best results for all settings. Adam (Kingma and Ba, 2014) is used as an optimizer. All experiments use H100 GPUs, with 7B models trained on one GPU and 13B models trained on two GPUs.

3.2 Translation Quality Results

Table 1 presents the results for the `Tower` series under an off-policy setting, measured by CometKiwi-XL and XCOMET. Except for closed-source models, all results are decoded by beam search with a beam size of 5. TowerInstruct-7B/-13B, and TowerInstruct-Mistral-7B are official implementations (Rei et al., 2024), supervised fine-tuned (SFT) on the corresponding base models using TowerBlock. We also conducted SFT on the Tower-Base series using 2K Best-of-N samples per direction, selected from our calibration dataset (§3.1) based on the highest CometKiwi-XXL scores. The resulting performance is comparable to the official instruction models.

When applying our calibration approach, very strong improvements can be observed across all directions, metrics, and base models. First, it leads to an average improvement of +2.8 points in KIWI-XL and +2.7 points in XCOMET over TowerInstruct-Mistral-7B. Additionally, Table 2 shows gains of +3.6 points in KIWI-XXL and +1.2 points in COMET, respectively. Second, this performance is comparable to that of the current top-performing system, that is Tower-70B-v2 equipped with 100-time-sampling MBR/TRR[3], while being approximately 200 times faster[4].

We also compare our approach with CPO (Xu et al., 2024), a widely used preference optimization method for translation. Following its original setup, we select the highest- and lowest-scoring candidates as accepted and rejected samples, respectively, and achieve consistent, substantial improvements over CPO.

3.3 Quality Estimation Results

As detailed in §2, we shared the objective for translation quality optimization and estimation, although supervisions are from machine annotations instead of human annotations. If optimized effectively, the resulting model should inherently acquire the ability to assess translation quality using

[2]Recent study (Wu et al., 2025a) shows that GPT-4o-mini can already serve as a strong translation system.

[3]TRR (Rei et al., 2024) denotes an ensemble strategy that applies reranking based on multiple metric model to select the best candidate from multiple sampled hypotheses. They report TRR results when it surpasses MBR.

[4]We roughly estimate the latency of the Tower-70B-v2 model to be 10 times that of the Tower-Mistral-7B model. Meanwhile, the former employs 100× sampling, while the latter uses beam search with a beam size of 5.

	Models	en→de KIWI-XL	en→de XCOMET	en→es KIWI-XL	en→es XCOMET	en→ru KIWI-XL	en→ru XCOMET	en→zh KIWI-XL	en→zh XCOMET	en→fr KIWI-XL	en→fr XCOMET
Closed	GPT-4o-mini	68.3	91.7	70.2	87.0	68.1	81.6	69.0	79.7	65.6	83.0
	GPT-4o	68.6	92.6	70.6	87.7	69.1	83.4	69.9	81.3	66.0	83.9
	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
	Tower-70B-v2 + MBR/TRR	72.3	–	74.5	–	74.2	–	72.6	–	–	–
	TowerInstruct-7B	69.0	91.7	70.8	86.9	69.0	81.5	68.5	78.7	67.9	84.1
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	70.0	92.0	70.8	86.5	69.6	81.6	68.4	77.9	68.0	83.7
	+ CPO	71.1	93.1	72.0	87.6	71.6	83.8	70.4	80.9	69.3	85.8
	+ Calibration (ours)	**71.6**	**93.6**	**73.5**	**89.0**	**72.4**	**84.8**	**70.4**	**81.0**	**70.0**	**86.8**
	TowerInstruct-13B	69.9	92.5	71.8	87.7	70.6	83.3	70.1	80.8	68.1	85.1
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	71.1	92.7	71.8	87.5	71.3	82.8	70.1	80.0	68.0	84.4
	+ CPO	70.5	92.2	72.0	87.7	71.9	84.0	70.3	81.4	68.8	85.5
	+ Calibration (ours)	**72.5**	**94.2**	**73.8**	**90.0**	**73.6**	**86.4**	**72.1**	**83.6**	**70.8**	**87.5**
	TowerInstruct-Mistral-7B	70.0	92.6	71.9	87.5	70.3	83.3	69.6	80.4	68.3	84.7
	+ SFT on BoN data	70.7	92.7	71.8	87.1	70.8	82.9	70.5	80.4	68.5	84.4
	+ CPO	71.2	93.0	73.1	89.0	72.3	85.1	71.8	83.6	70.0	86.9
	+ Calibration (ours)	**72.4**	**94.0**	**73.9**	**89.9**	**73.6**	**86.1**	**72.6**	**83.7**	**70.8**	**87.4**

	Models	en→nl KIWI-XL	en→nl XCOMET	en→it KIWI-XL	en→it XCOMET	en→pt KIWI-XL	en→pt XCOMET	en→ko KIWI-XL	en→ko XCOMET	Avg. KIWI-XL	Avg. XCOMET
Closed	GPT-4o-mini	69.4	88.9	68.1	83.7	71.2	87.6	73.2	84.2	69.2	85.3
	GPT-4o	70.6	90.5	68.7	85.7	71.5	88.5	73.7	85.6	69.8	86.6
	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
	Tower-70B-v2 + MBR/TRR	–	–	–	–	–	–	–	–	–	–
	TowerInstruct-7B	71.5	90.9	71.1	86.1	71.1	86.8	73.6	82.8	70.3	85.5
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	71.5	89.6	70.8	85.4	72.5	87.6	75.7	84.1	70.8	85.4
	+ CPO	71.9	90.9	72.2	86.7	73.4	88.7	76.1	87.2	72.0	87.2
	+ Calibration (ours)	**73.3**	**91.9**	**73.5**	**88.1**	**74.8**	**89.9**	**76.8**	**87.2**	**72.9**	**88.0**
	TowerInstruct-13B	71.7	91.0	71.1	87.3	72.1	88.2	75.4	84.8	71.2	86.7
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	71.7	90.4	71.6	86.1	73.0	88.1	76.2	85.2	71.6	86.4
	+ CPO	72.3	90.8	72.5	87.4	72.2	86.9	76.9	87.9	71.9	87.1
	+ Calibration (ours)	**73.9**	**92.6**	**73.9**	**89.3**	**75.2**	**90.4**	**78.0**	**89.5**	**73.8**	**89.3**
	TowerInstruct-Mistral-7B	71.9	91.1	71.6	87.2	72.1	88.0	74.2	85.6	71.1	86.7
	+ SFT on BoN data	72.3	90.7	71.6	86.2	72.7	87.9	76.2	86.0	71.7	86.5
	+ CPO	73.3	92.3	73.1	88.5	74.0	89.7	77.4	89.3	72.9	88.6
	+ Calibration (ours)	**74.2**	**93.2**	**74.1**	**89.6**	**75.1**	**90.7**	**78.1**	**89.7**	**73.9**	**89.4**

Table 1: en→xx translation qualities on WMT24 measured by CometKiwi-XL and XCOMET. Note that the Tower-v2 models, including Tower-70B-v2, have not been publicly released. We report their best results as published by Rei et al. (2024). For GPT-4o and GPT-4o-mini, we use the prompts following (Hendy et al., 2023). Results in other metrics can be found in Appendix A. Notably, according to Kocmi et al. (2024), improvements of ≥ 1.99 in XCOMET or ≥ 0.94 in COMET scores correspond to at least 90% estimated accuracy in human judgment—both of which are achieved by our method.

the hypothesis *log-likelihood* as a metric. In this section, we evaluate how effectively calibration can elicit this capability.

We use the WMT22 metric meta-evaluation dataset (Zerva et al., 2022) and follow the official practice to assess quality estimation ability using Spearman's and Kendall's correlation. We evaluate all training directions on Tower that overlap with the WMT dataset, namely, en→de and en→ru.

Figure 1 depicts the Spearman score (metric performance) and the corresponding translation performance under different settings for Tower-7B and Tower-13B, including: (1) supervised fine-tuning using varying amounts of best-of-N samples (400/800/1200/1600/2000 samples per direction), (2) scaling the base model size from 7B to 13B, and (3) applying our calibration method. It shows that:

(1) As more Best-of-N samples are included in SFT, translation performance progressively improves. Interestingly, the quality estimation ability (Spearman scores) increases from around 51.5 to 54.0 points. We attribute this to the fact that the model assigns higher likelihoods to better hypotheses. However, these improvements are limited and not general across languages, see Appendix B.

(2) Examining the effects of scaling, we observe that: (i) scaling up from 7B to 13B generally improves translation performance for both the original TowerInstruct models and the fine-tuned models; (ii) however, its impact on calibration, i.e., quality estimation ability, remains minimal.

(3) Our calibration method manifests very strong improvements in both translation and quality estimation. For example, when applying our method to TowerBase-13B, the resulting model surpasses some state-of-the-art systems in both translation performance and quality estimation ability, i.e., Tower-70B-v2+MBR/TRR and CometKiwi, at the same time.

Similar trends can be found in Figure 2 when we use the Kendall coefficient to measure the correlation. Results for en→ru are provided in Appendix B.

Overall, we observe a clear, albeit sometimes non-linear, correlation between the models' translation performance and their quality estimation ability. These results suggest—to some extent—a uni-

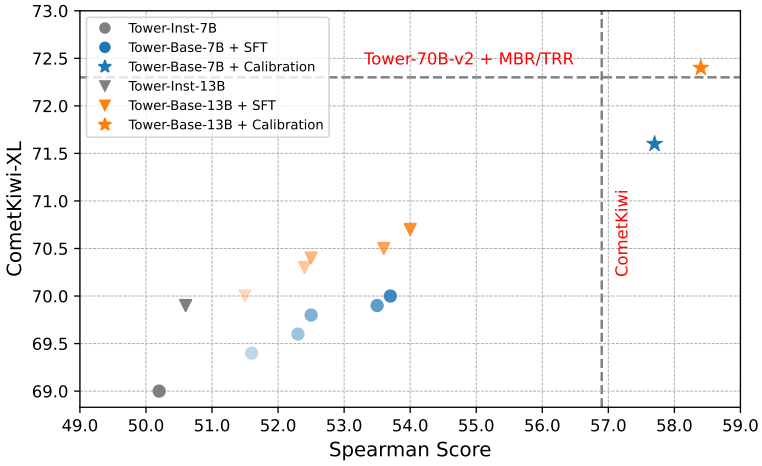


Figure 1: The Spearman coefficient and the corresponding translation performance in en→de direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ denotes the application of our calibration method, which simultaneously surpasses both the state-of-the-art translation system and the widely used quality estimation model.

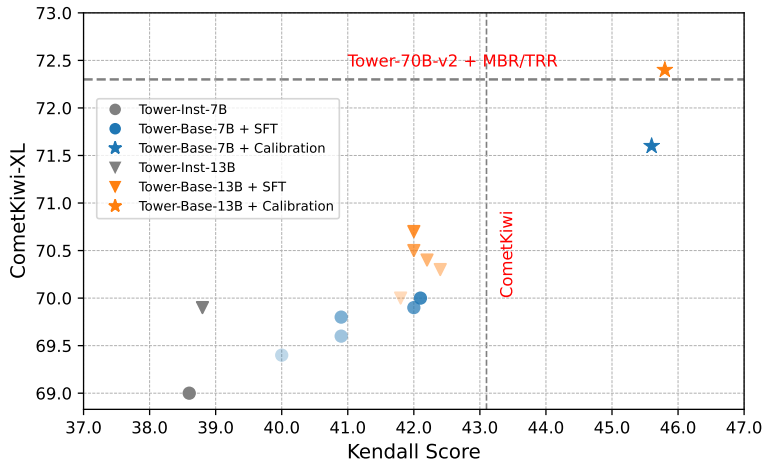


Figure 2: The Kendall coefficient and the corresponding translation performance in en→de direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ and ★ denote the application of our calibration method on 13B and 7B models, respectively.

fied perspective: a well-performing translation system should inherently 'know' what constitutes a good translation. In turn, we also suggest optimizing translation quality by improving calibration on LLMs, rather than relying solely on extreme scaling or supervised fine-tuning, as the latter approaches show relatively limited effectiveness.

4 Implementation of Our WMT25 Submission

In this section, we describe the implementation of our QE system for WMT25. To cover more languages, we use Gemma-3-12B as the base model instead of Tower, as noted in Section 3.1.

To construct the calibration dataset, similar to that in Section 3.1, we feed the source segments in the WMT25 general translation test set into GPT-4o-mini, using the prompts provided with the official test set[5], to generate 16 hypotheses per sample. Note that we use the WMT25 blind test sets rather than Flores. This choice is motivated by (1) better alignment with the domain used in testing and (2) consistency with WMT25's paragraph-level data format.

Following that in Section 3.1, the corresponding hypotheses are decoded using nucleus sampling with a top-p of 0.98 and a temperature of 1.0. Each sample is at the paragraph level, where "$\backslash n$" re-

[5]Official prompts can be found here.

mains in the original data as a separator. We also use CometKiwi-XXL to score each one-to-many translation pair in our synthetic dataset.

We apply the calibration method (Wu et al., 2025b), as we mentioned in Section 2, as the only post-training method on Gemma-3-12B.

Finally, the resulting model, trained on synthetic data derived from the WMT25 general translation test set, is used as a quality estimation model here. For each sample in the WMT25 QE dataset, we feed the source and target segments into our model (with the corresponding prompt) and directly use the **average log-likelihood** of the target segments as the quality assessment scores for submission.

For the performance of our system on the WMT25-QE track, please refer to this year's findings paper, which has not yet been officially released at the time of this submission.

5 On the Limitation of Using Uncertainty as a Proxy for Translation Quality

Although we demonstrate the effectiveness of using LLM uncertainty as a proxy for translation quality in Section 3.3, we also identify an important limitation—**this method will give significantly higher scores for translations that are from itself**—it favors its own output when it uses maximum a *posteriori* as the decoding rule.

This issue was not identified by (Wu et al., 2025b), as their QE testing set lacks such special conditions and does not include the model's own translation outputs during QE evaluation.

Meanwhile, we argue that this issue is likely to be a general limitation of most metrics based on translation uncertainty. More broadly, it to some extent aligns with observations about LLM-as-a-judge (Panickssery et al., 2024), where LLMs tend to favor their own generations.

Lastly, we anticipate that the official WMT25 QE test set will be particularly challenging for our metric. **As noted above, we use a single calibrated model to participate in both the general translation and quality estimation tasks at WMT25.** Therefore, if any hypotheses generated by our system appear in this year's QE test set, it is likely to assign them the highest scores. We have found indications of this, although we cannot confirm it because the official description of the test set has not yet been released.

References

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Di Wu, Seth Aycock, and Christof Monz. 2025a. Please translate again: Two simple experiments on whether human-like reasoning helps translation. *arXiv preprint arXiv:2506.04521*.

Di Wu, Yibin Lei, and Christof Monz. 2025b. Calibrating translation decoding with quality estimation on llms. *arXiv preprint arXiv:2504.19044*.

Di Wu, Shaomu Tan, Yan Meng, David Stap, and Christof Monz. 2024. How far can 100 samples go? unlocking zero-shot translation with tiny multi-parallel data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15092–15108, Bangkok, Thailand. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Results based on Tower in KIWI-XXL and COMET

Table 2 shows off-policy results measured by two other metrics, i.e., CometKiwi-XXL (abbreviated as KIWI-XXL) and COMET-22 (abbreviated as COMET). Very strong average performance improvements can be observed. For instance, +3.6 and +1.1 points of KIWI-XXL and COMET average gains are shown over TowerInstruct-Mistral-7B.

B Results on En→Ru Direction

In this appendix, we provide additional results in en→ru, complementing Section 3.3 of the main text.

Figure 3 shows the Spearman coefficient and the corresponding translation performance in the en→ru direction. Meanwhile, Figure 4 present the results using Kendall's τ for en→ru direction. It is clear that the main findings, as mentioned in Section 3.3, hold across language directions and statistics.

	Models	en→de		en→es		en→ru		en→zh		en→fr	
		KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET
Closed	GPT-4o-mini	76.4	82.7	76.3	83.8	75.5	82.5	75.8	84.6	74.7	81.5
Closed	GPT-4o	77.7	82.5	77.3	83.8	77.6	82.8	77.6	84.5	76.2	81.7
Closed	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
Closed	Tower-70B-v2 + MBR/TRR	–	–	–	–	–	–	–	–	–	–
	TowerInstruct-7B	76.5	81.2	76.3	82.8	75.9	81.1	74.8	83.1	76.7	81.2
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	77.2	81.3	75.8	82.4	76.2	80.9	74.4	82.4	76.2	81.0
	+ CPO	78.9	82.2	78.0	83.2	78.8	82.2	77.8	83.4	78.7	81.2
	+ Calibration	**79.5**	**82.8**	**79.8**	**83.7**	**80.4**	**82.9**	**78.0**	**83.2**	**80.2**	**81.7**
	TowerInstruct-13B	78.1	82.3	77.6	83.5	78.2	82.1	76.9	83.8	77.4	81.6
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	79.0	82.3	77.0	83.1	78.4	82.0	76.8	83.8	77.2	81.5
	+ CPO	79.1	82.1	78.6	82.5	80.3	82.6	78.0	83.4	79.3	81.5
	+ Calibration	**81.3**	**83.4**	**80.9**	**84.1**	**82.3**	**83.8**	**80.4**	**84.5**	**81.5**	**82.2**
	TowerInstruct-Mistral-7B	78.1	82.0	77.9	83.0	77.9	81.8	76.6	83.8	77.6	81.5
	+ SFT on BoN data	78.3	82.0	77.5	82.9	78.3	81.5	77.3	84.0	77.3	81.4
	+ CPO	79.6	82.2	79.9	83.3	80.5	82.7	79.7	84.8	79.9	81.8
	+ Calibration	**80.7**	**83.1**	**80.6**	**83.9**	**82.0**	**83.6**	**80.4**	**84.9**	**80.8**	**82.1**

	Models	en→nl		en→it		en→pt		en→ko		Avg.	
		KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET	KIWI-XXL	COMET
Closed	GPT-4o-mini	78.3	84.6	74.1	83.6	77.9	81.9	81.2	86.2	76.7	83.5
Closed	GPT-4o	80.7	84.6	76.0	83.8	79.1	81.9	82.3	86.2	78.3	83.5
Closed	Tower-70B-v2	–	–	–	–	–	–	–	–	–	–
Closed	Tower-70B-v2 + MBR/TRR	–	–	–	–	–	–	–	–	–	–
	TowerInstruct-7B	81.1	84.4	77.7	83.7	77.9	81.8	80.0	84.7	77.4	82.7
	TowerBase-7B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	80.5	83.5	76.9	83.4	78.6	81.5	82.3	85.3	77.6	82.4
	+ CPO	81.9	83.8	79.4	83.7	80.5	81.8	83.7	85.8	79.7	83.0
	+ Calibration	**83.6**	**84.8**	**81.0**	**84.3**	**81.9**	**82.7**	**84.6**	**86.1**	**81.0**	**83.6**
	TowerInstruct-13B	81.4	84.6	78.4	84.2	79.1	82.5	82.9	85.5	78.9	83.4
	TowerBase-13B	–	–	–	–	–	–	–	–	–	–
	+ SFT on BoN data	80.8	84.3	77.9	83.8	79.5	81.7	83.6	85.7	78.9	83.1
	+ CPO	82.5	84.2	80.2	83.8	79.2	80.7	85.0	86.5	80.2	83.0
	+ Calibration	**84.5**	**85.1**	**82.1**	**84.6**	**82.8**	**82.7**	**86.2**	**87.1**	**82.4**	**84.2**
	TowerInstruct-Mistral-7B	81.5	84.6	79.0	84.0	79.3	82.2	81.7	85.3	78.8	83.1
	+ SFT on BoN data	81.4	84.2	78.4	83.7	79.6	81.7	83.8	86.1	79.1	83.0
	+ CPO	83.9	84.8	80.7	84.0	81.9	82.2	85.9	86.9	81.3	83.6
	+ Calibration	**84.6**	**85.2**	**82.3**	**84.8**	**83.2**	**83.0**	**86.9**	**87.3**	**82.4**	**84.2**

Table 2: Evaluation of en→xx translation on WMT24 using CometKiwi-XXL and COMET. Results are reported for all languages covered during Tower-v1 pretraining. Note that the Tower-v2 models, including Tower-70B-v2, have not been publicly released. For GPT-4o and GPT-4o-mini, we use the prompts following (Hendy et al., 2023). Notably, according to Kocmi et al. (2024), improvements of ≥ 1.99 in XCOMET or ≥ 0.94 in COMET scores correspond to at least 90% estimated accuracy in human judgment — both of which are achieved by our method.

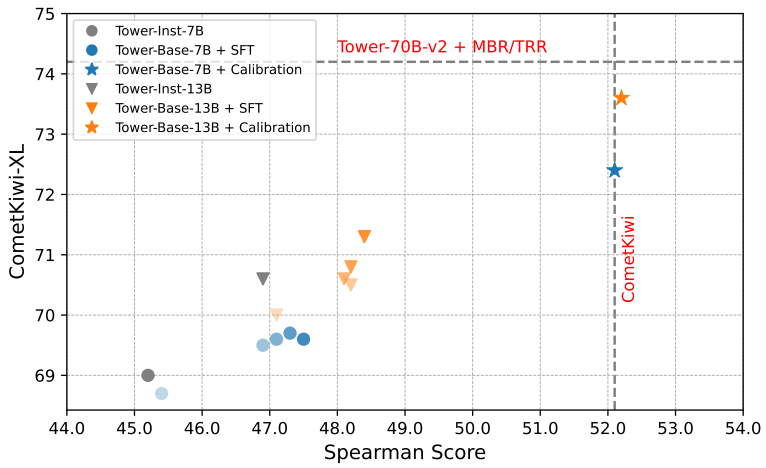


Figure 3: The Spearman coefficient and the corresponding translation performance in en→ru direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ and ★ denote the application of our calibration method on 13B and 7B models, respectively.

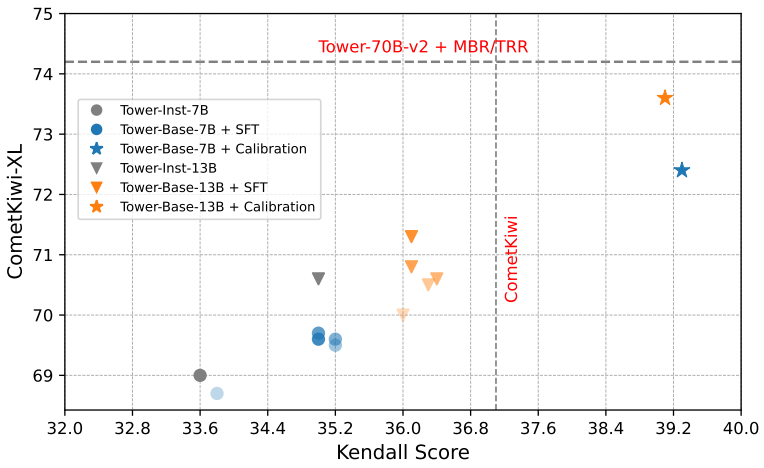


Figure 4: The Kendall coefficient and the corresponding translation performance in en→ru direction under different settings for the Tower series models. The color gradients of ▼ and ●, from lighter to darker shades, indicate the results of fine-tuning with varying amounts of Best-of-N data, from 400 to 2000 samples. ★ and ★ denote the application of our calibration method on 13B and 7B models, respectively.

Can QE-informed (Re)Translation lead to Error Correction?

Govardhan Padmanabhan

Institute for People-Centred AI
University of Surrey, United Kingdom
gp00816@surrey.ac.uk

Abstract

The paper presents two approaches submitted to the WMT 2025 Automated Translation Quality Evaluation Systems Task 3 - Quality Estimation (QE)-informed Segment-level Error Correction. While jointly training QE systems with Automatic Post-Editing (APE) has shown improved performance for both tasks, APE systems are still known to *overcorrect* the output of Machine Translation (MT), leading to a degradation in performance. We investigate a simple training-free approach - QE-informed Retranslation, and compare it with another within the same training-free paradigm. Our winning approach selects the highest-quality translation from multiple candidates generated by different LLMs. The second approach, more akin to APE, instructs an LLM to replace error substrings as specified in the provided QE explanation(s). A conditional heuristic was employed to minimise the number of edits, with the aim of maximising the Gain-to-Edit ratio. The two proposed approaches achieved a ΔCOMET score of 0.0201 and -0.0108, respectively, leading the first approach to achieve the winning position on the subtask leaderboard.

1 Introduction

Large Language Models (LLMs) have advanced the field of Machine Translation (MT), given their support for longer input context length and ability to generate text in a natural tone. However, translation quality is still limited for languages other than English and for domain-specific translations (Fernandes et al., 2025). Evaluating MT output for quality is critical to understand the reliability and suitability of translation systems, and more importantly, to be able to perform accurate corrections. The WMT24 Metrics Shared Task found that neural-based learned metrics like COMET or xCOMET were superior when evaluating LLM-generated translations, compared to the traditional statistical metrics like BLEU, or chrF (Freitag et al., 2024).

As LLMs are typically trained on a large general dataset, performance in domain-specific MT can often fall short as they may not properly render key terminologies or stylistic conventions. As such, Automatic Post-Editing (APE) is vital to fixing MT errors. However, they are prone to overcorrecting. One such example of mitigating over-correcting, proposed by Deoghare et al. (2025), is to utilize word-level Quality Estimation (QE) to limit edits only on the specified error segments. Despite recent efforts to reduce over-correction (Deoghare et al., 2023, 2024), APE models still fall short of the required semantically coherent output. Therefore, we ask the titular question - "Can QE-informed re-translation help overcome MT errors?", and discuss two approaches as a comparative evaluation.

This paper describes two participation systems, both utilizing pre-trained and open-source LLMs, for the WMT 2025 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction. The primary approach leverages multiple LLMs for MT and selects the best output using QE. In the secondary approach, an LLM is prompted to replace error-segments in the provided MT. These error segments are identified by the explainable QE provided within the dataset.

2 Related Work

Quality Estimation (QE) QE is an automated evaluation framework that predicts a score indicating whether the translation is good or not (Yvon, 2019). COMET (Cross-Lingual Optimized Metric for Evaluation of Translation) is an automatic metric to evaluate the quality of machine translation using deep learning models (Rei et al.,

Proceedings of the Tenth Conference on Machine Translation, pages 984–993
November 8-9, 2025

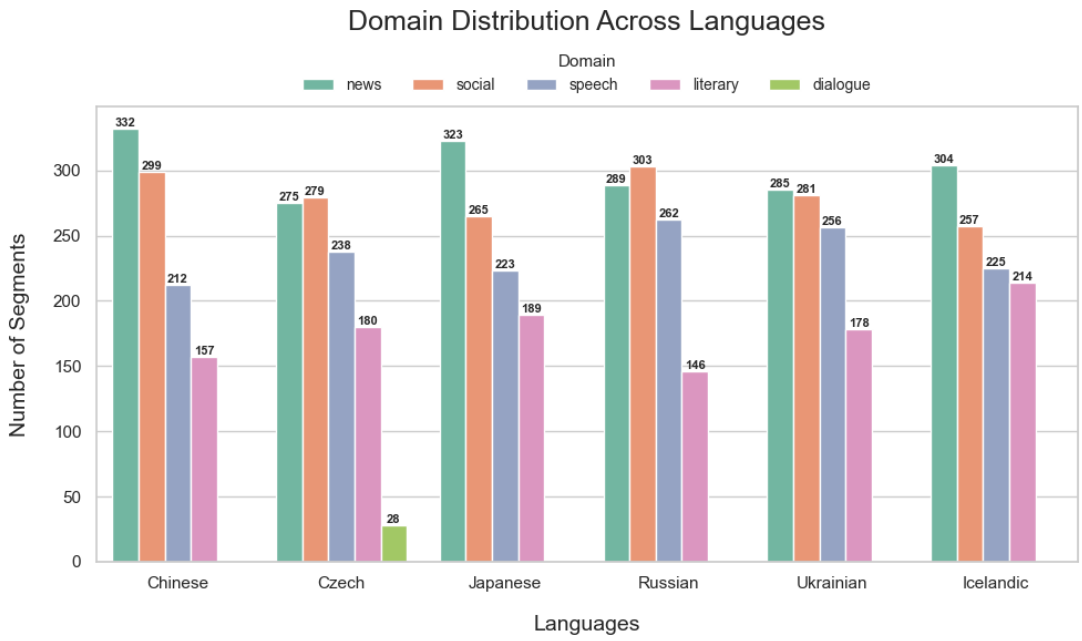


Figure 1: Domain distribution within each language

2020). COMETKiwi is a hybrid machine translation quality estimation (MTQE) model that combines COMET and OpenKiwi (Rei et al., 2022). It achieved top performance in the WMT 2022 shared task and has since been widely adopted as a state-of-the-art benchmark in MTQE.

Automatic Post-Editing (APE) APE is an automated system that corrects MT output, without human involvement (do Carmo et al., 2021). Chatterjee et al. (2018) describe combining QE and APE in three ways: using QE as an APE activator when the MT output is poor, as guidance to help the APE decoder decide which tokens to change, and as a selector to choose between the raw MT and post-edited output.

WMT24 QE-APE The previous year's WMT competition focused on sentence-level quality estimation and error span predictions (Zerva et al., 2024). QE was further incorporated into APE. The dataset for the QE-APE primarily consisted of English–Hindi (En-Hi) and English–Tamil (En-Ta) pairs. The source (SRC) English sentence, the target (TGT) translation provided by an unspecified neural machine translator, and a human post-edit (PE) version of the translation made by native speakers were included to support both quality estimation and automatic post-editing tasks.

The HW-TSC team (Yu et al., 2024) utilized `Llama3-8B-Instruct` for En-Hi pairs. The model first underwent continual pretraining using low-rank adaptation (LoRA) on SRC and TGT data, and was further supervised fine-tuned on the PE data with a custom prompt. For En-Ta pairs, they trained a custom transformer then performed APE fine-tuning. This system achieved 0.851 and 0.918 COMET scores for En-Hi and En-Ta translation tasks, respectively.

The IT-Unbabel team utilized `xTower` for generating corrected translations, along with a quality estimation model to decide whether to use the original translation or the `xTower` output. This approach achieved 0.8646 COMET score for En-Hi pairs, and 0.9163 for En-Ta pairs.

IT-Unbabel's solution of utilizing an LLM along with QE as a selector served as the inspiration for the primary approach

3 Methodology

3.1 Dataset Analysis

The complete test data provided with the task was exclusively used. This dataset consists of 6,000 machine translations from English to Chinese, Czech, Japanese, Icelandic, Russian, and Ukrainian, with 1,000 instances for each language pair. The texts cover a range of domains, specifically news, social, speech, literary, and dialogue. As shown in Figure 1, the domains are not evenly distributed. The *news* domain is the most prominent overall with 1,808 entries, while the *dialogue*

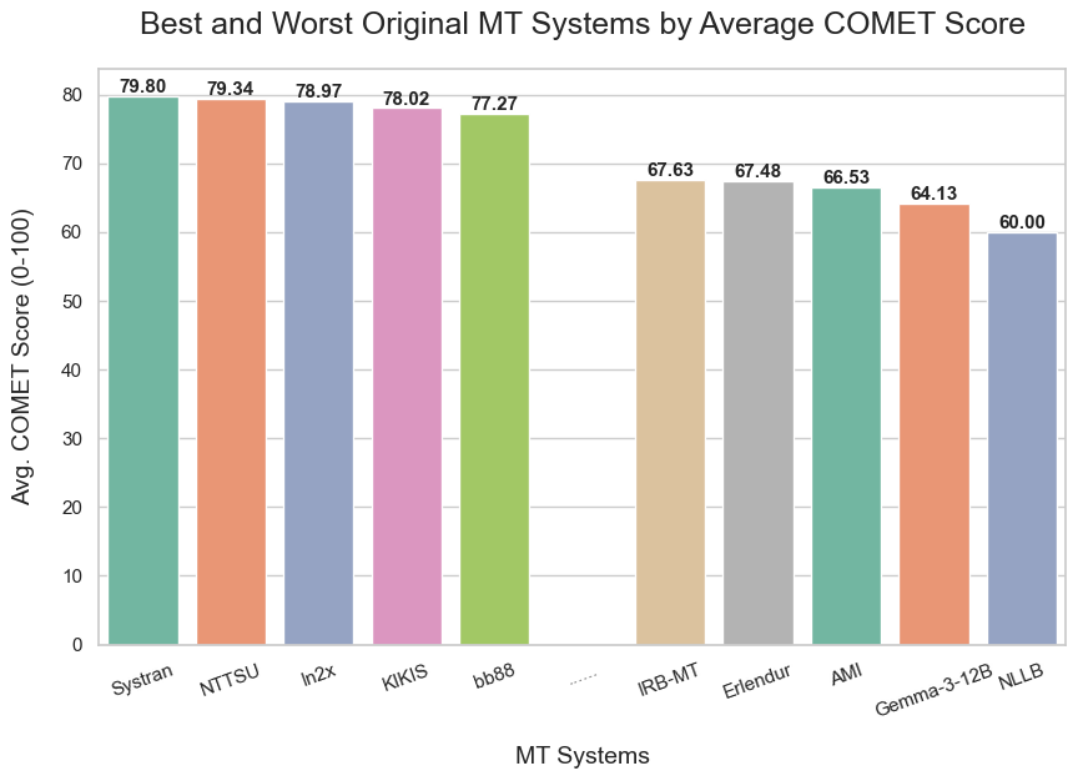


Figure 2: System performance in hypothesis_segment

domain has the lowest representation with 28 segments, all appearing only in the English-to-Czech group.

The original translations under the "hypothesis_segment" column were generated by 38 different translation systems, including LLMs like `GPT4.1`, `Claude-4`, `DeepSeek-v3`, and more. Figure 2 shows the five best and worst systems according to their average translation quality scores.

3.2 Approaches

3.2.1 Primary Approach - "Best MT Wins"

In this approach, multiple LLMs were used to translate the English texts from scratch, without additional information or context. The resulting candidate translations were then evaluated with the *wmt22-cometkiwi-da* model, which provided QE scores based on the source English text and translated system outputs. The translation with the highest QE score was selected as the final output. By re-framing the APE task as re-translation, QE serves as a selector in place of traditional decoders. Similar approaches have also been explored, where QE is used in multi-hypothesis selection (Yu et al., 2024; Laki and Yang, 2018; Lu and Zhang, 2019), further supporting the view of QE as a decision mechanism in MT improvement. This approach therefore functions as a systematic probe of QE-based re-ranking, illustrating both its potential to establish an empirical upper bound on metric-based performance and its limitations in terms of computational cost and efficiency.

`Aya-Expanse-8B`, `GPT-SW3-6.7B`, `Tri-7B`, `GLM4-9B`, `Phi4-mini-instruct`, and `TowerPlus-9B` models were used. These models were selected because of their reported performance, robustness, popularity, and recency. Models like `GPT-SW3-6.7B`, `Tri-7B`, and `GLM4-9B` were selected owing to their specialized training in certain languages, specifically Icelandic, Japanese, and Chinese respectively.

The prompts for `Tri-7B`, `Phi4-mini-instruct`, and `TowerPlus-9B` are reminiscent of their translation prompts available in their Huggingface model cards, being variations of:

> Translate from English to *{Language}*

System prompts for the other models were more involved, as using the same prompt resulted in some hallucinations, chain of thought, or worse translation quality during limited internal testing.

For `Aya-Expanse-8B`:

> You are a helpful bilingual assistant that correctly translates the user's input text from English to *{Language}*.
> When translating, you must use the same tone and intent of the English text. You will include any and all special characters from the input.
> If there is no proper translation for an English word or phrase, you can use the English word or phrase in place.

For `GPT-SW3-6.7B`:

> <|endoftext|><s>
> System:
> You are a bilingual assistant that objectively translates the user's input text from English to Icelandic.
> You will include any and all special characters from the input.
> If there is no proper translation for an English word or phrase, you can use the English word or phrase in place.
> <s>
> User:
> *{original English text}*
> <s>
> bot:

For `GLM4-9B`:

> You are a bilingual assistant that correctly translates the user's input text from English to Chinese.
> When translating, you must use the same tone and intent of the English text. You will include any and all special characters from the input.
> If there is no proper translation for an English word or phrase, you can use the English word or phrase in place.
> Output only the translation!

Except `TowerPlus-9B`, the remaining models do not support all required languages. Hence, they translated only their supported language(s), and the rest were omitted. After all models were successfully executed, the QE score via COMETKiwi between all translations, including the original systems', was used to determine which MT to use.

3.2.2 Secondary Approach - "Fill in the Blanks"

The provided test data includes *error spans* for the translations. Using fine-grained QE signals to guide targeted corrections, this approach investigates whether restricting edits to QE-highlighted segments could yield improvements with fewer changes, thereby increasing both efficiency and interpretability compared to full re-translation.

Using these error spans, the corresponding substring(s) in translation is replaced with a "**__BLANK__**" token, emulating a multilingual masked language modeling task. The text domain is also used to provide additional context. An example is included in the prompt to guide the model's behavior, serving as a one-shot example. Different examples were used in the prompt for different languages.

This approach utilizes `TowerPlus-9B`, an open source LLM with 9 billion parameters based on Gemma2. This model was selected because of its training for translation-related tasks along with instruction tuning, a context window of 8192 tokens, and has support of the languages of this task: English, Chinese, Czech, Japanese, Icelandic, Russian, Ukrainian. It has also been shown to outperform larger parameter models like `Gemma2-27B` and `Llama3.3-70B` on translation performance (Rei et al., 2025). The 9B variant was specifically selected due to resource and time constraints.

Provided below is the system prompt template for Russian language translations:

> You are a helpful assistant that corrects a Russian translation by filling in the blanks. Use the English sentence for context. Complete the task while maintaining the tone of a *{domain}*.
> Important - Do not use any of the specified wrong words. Replace each __BLANK__ token with an appropriate word or phrase that matches the original meaning, tone, and context of the English sentence.
>
> Example (social domain):

Russian with __BLANK__: Многие молодые люди сегодня предпочитают __BLANK__ в кофейнях, а не дома.
English: Many young people today prefer to hang out in coffee shops rather than at home.
Wrong translation: Многие молодые люди сегодня предпочитают учиться в кофейнях, а не дома.
Wrong words: [' учиться']
Corrected Russian sentence: Многие молодые люди сегодня предпочитают проводить время в кофейнях, а не дома.

Russian with __BLANK__: *{mt with __BLANK__}*
English: *{source text}*
Wrong translation: *{mt text}*
Wrong words: *{list of substrings removed}*
Corrected Russian sentence:

To increase translation quality while keeping changes minimal (gain-to-edit ratio), a conditional masking heuristic based on error severity and overall QE score was employed. The pseudocode for this method is provided in Algorithm 1.

Algorithm 1 Conditional masking based on severity and score

```
Original QE Score: Float x
if x >= 0.90 then
    Proceed without masking
else if x > 0.50 then
    if only minor severity error spans then
        Mask minor error spans
    else
        Mask all non-minor error spans
    end if
else
    Mask all error spans
end if
```

4 Results and Discussion

Tables 1 and 3 present language-wise results for the primary and secondary approaches, respectively. The WMT'25 shared task tracks two main metrics: ΔCOMET and Gain-to-Edit Ratio (referred to as "G2E Ratio"), with the overall ΔCOMET score serving as the primary selection criterion. Here, ΔCOMET is computed as the difference between the COMET scores of the proposed approach and the baseline translation. The G2E Ratio is calculated as ΔCOMET divided by the total edit rate. BLEU and chrF++ scores are also included in the tables as supplementary information, but they are not the focus of the analysis. The best-performing ΔCOMET and G2E Ratio scores across the two tables are highlighted in bold.

4.1 Primary Approach - "Best MT Wins"

Language	ΔCOMET	G2E Ratio	BLEU	chrF++
Icelandic	$3.65e-2$	$1.27e-3$	69.24	78.37
Russian	$2.01e-2$	$7.10e-4$	68.56	79.15
Czech	$1.89e-2$	$8.15e-4$	73.85	82.29
Chinese	$1.84e-2$	$1.64e-4$	39.35	58.41
Ukrainian	$1.63e-2$	$6.36e-4$	71.36	80.85
Japanese	$1.03e-2$	$1.41e-4$	54.33	65.44
Average	$\mathbf{2.01e-2}$	$\mathbf{6.22e-4}$	62.78	74.08

Table 1: Language-wise Results for Primary Approach

Table 1 shows that an ensemble of diverse models helps significantly improve the translations, especially in low-resource languages like Icelandic with roughly 3% improvement.

As this approach selects the translation with the best QE score, not all systems or model responses contributed equally to the final output. As shown in Table 2, `TowerPlus-9B` and the original translations have contributed the most, while `Tri-7B` did not contribute at all. Despite `GPT-SW3-6.7B`, `Tri-7B`, and `GLM4-9B` trained primarily for use in Nordic and Altaic languages and Chinese, other models provided better translations. Figure 3 shows how different models contributed to the final response in each language.

System	Contribution
Tower+ 9B	2836
Original	2829
GLM4 9B	149
Phi4 Mini Instruct	106
Aya Expanse 8B	76
GPT-SW3 6.7B v2 Instruct	9
Tri 7B	0

Table 2: System-wise Contribution to the Output

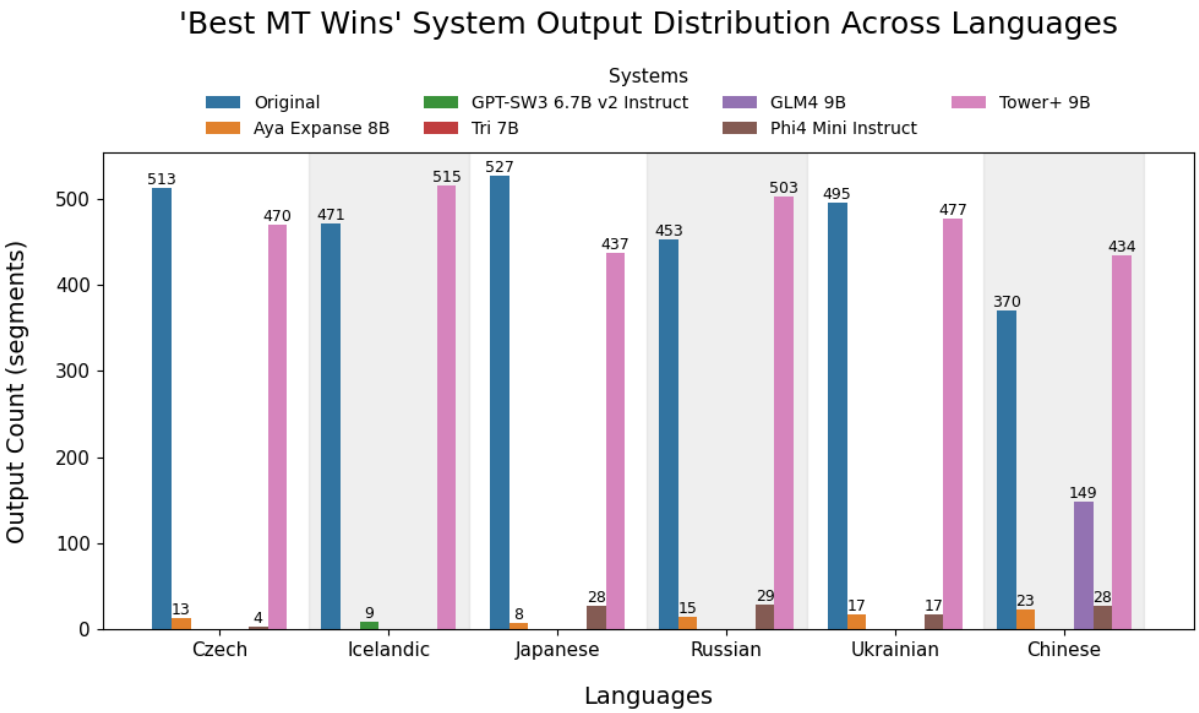


Figure 3: Language-wise count of model responses used in primary approach's final output

The example in Appendix A is an instance where the direct translation by `TowerPlus-9B` was the best-performing. Interestingly, the corresponding output from the secondary approach is of lower quality. This demonstrates the powerful translation capabilities of the model and the need for better prompt engineering in the secondary approach.

4.2 Secondary Approach - "Fill in the Blanks"

Language	ΔCOMET	G2E Ratio	BLEU	chrF++
Czech	$-7.24e-3$	$-5.80e-7$	92.37	95.14
Russian	$-8.03e-3$	$-6.00e-7$	92.56	95.31
Icelandic	$-1.00e-2$	$-8.20e-7$	74.27	82.18
Chinese	$-1.30e-2$	$-4.83e-5$	29.52	83.59
Japanese	$-1.34e-2$	$-4.44e-5$	25.77	82.59
Ukrainian	$-1.35e-2$	$-1.26e-6$	91.77	94.43
Average	$-1.08e-2$	$-1.59e-5$	67.71	88.87

Table 3: Language-wise Results for Secondary Approach

Table 3 shows that this approach does not provide much benefit to the overall translation task, with roughly -0.7% to -1.4% quality degradations.

Appendix C is an example where this approach yields a better translation than the original. The original Czech translation by `DeepSeek-v3` has 3 major errors and 1 minor error. This improvement by the model could be owed to the fact that, on top of its multilingual capabilities, `TowerPlus-9B` was fine-tuned on instruction data from different models, including `DeepSeek-v3`.

There are possible factors why this approach did not perform well: use of bigger and closed-source models in the original MT, use of varied systems and models, inclusion of low-resource languages, correcting only critical + major error spans and ignoring minor error spans in some instances, and more. In the Appendix B example, the original system used Massively Multilingual Neural Machine Translation (MMMT). Although more capable models have been released since then, including `TowerPlus-9B`, output artifacts ("__HEARTBREAK__") and leakage ("Corrected words: [...") affect the translation's quality. This shows that further prompt tuning and output post-processing is needed.

5 Conclusion and Future Work

This paper described two translation approaches submitted to the WMT 2025 QE-informed Segment-level Error Correction task. The first approach used QE as a selector among multiple

LLM translation outputs, resulting in an overall ΔCOMET of 0.0201. The second approach utilized `TowerPlus-9B` exclusively to replace erroneous words in the MT by masking substrings highlighted in the error spans with a blank token, resulting in ΔCOMET of -0.0108. Custom prompts were designed to instruct the model in correcting the translation, similar to a fill-in-the-blank task.

While the first approach showed positive improvements by using QE as a selector, it ultimately depends on the model and system selection. Further exploration of system combinations could yield better performance. The second approach, which performed worse, corrected translations by filling in error segments in the MT. As future work, a two-model system could be explored: a smaller LM to suggest words or phrases for masked tokens via masked language modeling, and a larger LLM to select the most suitable ones to produce a higher-quality translation.

Limitations

Due to limited time and compute resources, the overall experimental design favored n-shot prompting with LLMs for their ease of use and availability of pre-trained weights. Additionally, the model selection was guided by convenience and practical factors such as parameter count, recency, and language compatibility or specificity. These limitations also limited the scope for testing and prompt optimization.

Though "Best MT Wins" reframes APE as retranslation through QE-based selection, this approach is impractical as it requires generating full hypotheses from large models. While the direct translation capabilities of `TowerPlus-9B` even slightly surpassed the original translation system, the other 5 LLMs used were less effective in comparison, resulting in inefficient use of and wasted time and computational resources.

For the "Fill in the Blanks" approach, a lack of proper prompt tuning and post-processing degraded the output quality, indicating that prompt engineering and output handling are important when using LLMs for specific tasks.

More broadly, both approaches rely exclusively on automatic QE metrics such as COMET, which, while effective for shared task evaluation, are primarily trained on English-centric data. The absence of human evaluation limits the ability to validate whether metric-based gains reflect true improvements in translation quality, especially for non-English language pairs.

6 Acknowledgements

My sincerest gratitude to Dr Diptesh Kanojia, Archchana Sindhujan and Sourabh Deoghare from the shared task team for their valuable feedback and guidance, encouragement, paper review, and assistance with LaTeX formatting.

References

Rajen Chatterjee, Matteo Negri, Marco Turchi, Frédéric Blain, and Lucia Specia. 2018. Combining quality estimation and automatic post-editing to enhance machine translation output. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 26–38, Boston, MA. Association for Machine Translation in the Americas.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2024. Together we can: Multilingual automatic post-editing for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10800–10812, Miami, Florida, USA. Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, and Pushpak Bhattacharyya. 2025. Giving the old a fresh spin: Quality estimation-assisted constrained decoding for automatic post-editing.

Sourabh Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe, and Pushpak Bhattacharyya. 2023. Quality estimation-assisted automatic post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1686–1698, Singapore. Association for Computational Linguistics.

F. do Carmo, Dimitar Shterionov, Joss Moorkens, Andy Way, Federico Gaspari, and Joachim Wagner. 2021. A review of the state-of-the-art in automatic post-editing. *Machine Translation*, 35:101–143.

Patrick Fernandes, Sweta Agrawal, Emmanouil Zaranis, André F. T. Martins, and Graham Neubig. 2025. Do llms understand your translations? evaluating paragraph-level mt with question answering. *Preprint*, arXiv:2504.07583.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami,

Florida, USA. Association for Computational Linguistics.

László János Laki and Zijian Győző Yang. 2018. Combining machine translation systems with quality estimation. In *Computational Linguistics and Intelligent Text Processing*, pages 435–444, Cham. Springer International Publishing.

Jinliang Lu and Jiajun Zhang. 2019. Select the best translation from different systems without reference. In *Natural Language Processing and Chinese Computing*, pages 355–366, Cham. Springer International Publishing.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jiawei Yu, Xiaofeng Zhao, Min Zhang, Zhao Yanqing, Yuang Li, Su Chang, Xiaosong Qiao, Ma Miaomiao, and Hao Yang. 2024. HW-TSC's participation in the WMT 2024 QEAPE task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 541–546, Miami, Florida, USA. Association for Computational Linguistics.

François Yvon. 2019. Quality estimation for machine translation. *Computational Linguistics*, 45(2):391–394.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Example of Primary Approach

Language: Chinese

Original System: Wenyiil

Original English Text:
A brief note on Play Station controllers to round out the 1st party lineup: I never liked them.\nDual Shock 1-3 felt exactly the same to me with exactly the same flaws: Oversized thumbsticks placed too close together, mushy face/shoulder/trigger buttons that seemed to wear out very quickly, and a D-pad that just felt wrong. Dual Shock 4 fixed a lot of problems, but introduced a bunch of new ones. I got a Dual Sense because some friends had a hard time with asymmetrical thumbsticks, and for some reason local stores were selling them for $20 less than DS4 controllers.\nI don't really use this controller: Face buttons are spaced too far apart, my audio in/output keeps defaulting to the controller when it's connected, the bumpers are weirdly tall, and I keep accidentally pressing the track pad when trying to hit menu/share.\nThey are fine guest controllers that have come a long way in terms of design, but for me they're a novelty at best.

Original Chinese MT:
于PlayStation手柄，里提一下，算是第一方手柄容的一个充：我从来都不喜它。\n在我看来，DualShock 1到3代手感完全一，缺点也如出一：杆帽太大且位置太近，面板按、肩和扳机手感很肉，而且似乎特容易磨，十字的感也是不。DualShock 4修正了多，但也引出了一堆新毛病。我了一个DualSense手柄，是因有些朋友用不非称杆，并且出于某原因，本地商店里它的售价比DS4手柄要便宜20美元。\n我其不怎用个手柄：面板按距太，次接手柄的音入/出都会自切去，肩的高度很奇怪，而且当我想按菜/分享会触到触摸板。\n它作客人用的手柄不，在上也确有了足的，但我而言，充其量算是个新奇玩意儿。

New MT:

于 PlayStation 手柄，我想充一点：我从来不喜它。\nDual Shock 1-3 我来感完全一，而且有同的缺点：拇指杆太大，距离太近；部/肩部/扳机按的，似乎很快就会磨；D 方向感不。Dual Shock 4 解决了好多，但也引入了好多新。我了一个 Dual Sense，因有些朋友不称的拇指杆感到不适，而且不知什原因，当地商店的价格比 DS4 手柄便宜 20 美元。\n我很少用个手柄：部按之的距离太，接我的音入/出是默置手柄；冲器很奇怪地高；当我点菜/分享，我是不小心按下了触控板。\n它是不的客用手柄，在上已有了很大的改，但我来，它充其量只是新奇玩意儿。

Original COMET Score: 0.7751

New COMET Score: 0.7797

Approach 1's Edited COMET Score: 0.7723

B Negative Example for Secondary Approach

Language: Japanese

Original System: MMMT

Original English Text:
Kyle nodded, the corners of his mouth twitching as he fought back a smile. "Is that right? Sorry to let you down. Guess it's a good thing I hung up the skates when I did, spare a few more fantasy teams the heartbreak."\nDeclan opened his mouth, then quickly shut it again, heat creeping up the back of his neck. "I mean…your Cup run with the Kings was insane, though. That game seven OT goal? Unreal." He scratched the back of his neck. "So… uh, what are you doing at St. Cassian's? Are you coaching here now? I didn't see you on the website."

Original Japanese MT:
カイルは頷き、微笑みを抑えながら口角をそらせた。「そうですか。がっかりさせちゃってごめんなさい。急いでスケートを引退しておいてよかったと思うのは、他のファンタジーチームがまた心を痛めることを避けるためでしょう。」\nデクランは口を開けてすぐに閉じ、首の後ろに熱さがこみ上げてきた。"いや…キングスとのカップ戦は、ただただすごいね。あの第7戦の延長戦のゴールは…信じられないよ。"彼は首の後ろを掻いた。「それで…あんた、セント・カシアンズでは何をしてるんだ？今はここでコーチをやっているのか？ウェブサイトにはあなたがいる姿が見当たらなかったけど。」

Modified Japanese MT:
カイルは頷き、微笑みを抑えながら口角をそらせた。「そうですか。がっかりさせちゃってごめんなさい。急いでスケートを引退しておいてよかったと思うのは、他のファンタジーチームがまた__HEARTBREAK__を痛めることを避けるためでしょう。」\nデクランは口を開けてすぐに閉じ、首の後ろに熱さがこみ上げてきた。"いや…キングスとのカップ戦は、ただただすごいね。あの第7戦の延長戦のゴールは…信じられないよ。"彼は首の後ろを掻いた。「それで…あんた、セント・カシアンズでは何をしてるんだ？今はここでコーチをやっているのか？ウェブサイトにはあなたがいる姿が見当たらなかったけど。」
Corrected words: ['HEARTBREAK', 'カップ戦']

Original COMET Score: 0.8183

New COMET Score: 0.7314

Error Spans:
$\{'start\_i' : 87, 'end\_i' : 88, 'severity' : 'major'\}$,
$\{'start\_i' : 150, 'end\_i' : 154, 'severity' : 'major'\}$

C Positive Example for Secondary Approach

Language: Czech

Original System: DeepSeek-v3

Original English Text: The duchess and her husband, Prince Harry, have to do something and good luck to them, but it's hard not to study the rollout of As Ever for signs of the widening gap between Meghan's self-image and how the rest of the world sees her. ""As ever,""writes the duchess on Instagram, "means "as it's always been"or some even say "in the same way as always,""the "some"in this sentence apparently referring to the dictionary definition of a two-word phrase that no one has ever had trouble understanding.

Original Czech MT: Vévodkyně a její manžel, princ Harry, musí něco dělat, a přejme jim hodně štěstí, ale nelze si nevšimnout, že uvedení značky As Ever odhaluje rostoucí propast mezi Meghaniným sebepojetím a tím, jak ji vnímá zbytek světa. „‚As ever‘," píše vévodkyně na Instagramu, „znamená ‚jak to vždy bylo‘, nebo někteří dokonce říkají ‚stejným způsobem jako vždy‘." Tito „někteří" v této větě zřejmě odkazují na slovníkovou definici dvouslovného výrazu, kterému nikdo nikdy nerozuměl.

Modified Czech MT: Vévodkyně a její manžel, princ Harry, musí něco udělat, a přejme jim hodně štěstí, ale nelze si nevšimnout, že uvedení značky As Ever odhaluje rostoucí propast mezi Meghaniným sebepojetím a tím, jak ji vnímá zbytek světa. „‚As ever‘," píše vévodkyně na Instagramu, „znamená ‚jak to vždy bylo‘, nebo někteří dokonce říkají ‚stejným způsobem jako vždy‘." Tito „někteří" v této větě zřejmě odkazují na slovníkovou definici dvouslovného výrazu, kterému nikdo nikdy neměl problém porozumět.

Original COMET Score: 0.8215

New COMET Score: 0.8317

Error Spans:
$\{'start\_i' : 42,' end\_i' : 53,' severity' : 'major'\}$,
$\{'start\_i' : 174,' end\_i' : 180,' severity' : 'minor'\}$,
$\{'start\_i' : 447,' end\_i' : 467,' severity' : 'major'\}$,
$\{'start\_i' : 468,' end\_i' : 469,' severity' : 'major'\}$

RankedCOMET: Elevating a 2022 Baseline to a Top-5 Finish in the WMT 2025 QE Task

Sujal Maharjan[1*], **Astha Shrestha**[1*]
[1] IIMS College, Kathmandu, Nepal
{sujalmaharjan007, aasthashrestha688}@gmail.com
*These authors contributed equally to this work

Abstract

This paper presents **rankedCOMET**, a lightweight per-language-pair calibration applied to the publicly available Unbabel/wmt22-comet-da model that yields a competitive Quality Estimation (QE) system for the WMT 2025 shared task. This approach transforms raw model outputs into per-language average ranks and min–max normalizes those ranks to $[0, 1]$, maintaining intra-language ordering while generating consistent numeric ranges across language pairs. Applied to 742,740 test segments and submitted to Codabench, this unsupervised post-processing enhanced the aggregated Pearson correlation on the preliminary snapshot and led to a 5th-place finish. We provide detailed pseudocode, ablations (including a negative ensemble attempt), and a reproducible analysis pipeline providing Pearson, Spearman, and Kendall correlations with bootstrap confidence intervals.

1 Introduction

Machine translation Quality Estimation (QE) predicts the quality of a translation without reference texts. For many production environments, affordable and reliable QE is more valuable than marginal benefits from retraining large models. We therefore analyze whether a robust, publicly available metric model (Unbabel/wmt22-comet-da; Rei et al., 2022) can maintain its competitiveness in 2025 when paired with a simple, computationally inexpensive post-processing step.

Our per-language rank-based calibration—**rankedCOMET**—is model-agnostic, computationally efficient, and empirically effective: it improved aggregated Pearson correlation on the preliminary Codabench verification snapshot and was adequate to reach 5th place on that snapshot. To evaluate our approach, we benchmark it against alternative calibration techniques and offer a suite of diagnostics so others can replicate and extend our findings.

2 Related Work

Neural evaluation metrics (COMET and follow-ups) are widely used for MT evaluation (Rei et al., 2020, 2022). Calibration techniques (Platt scaling, isotonic regression) are standard in classification/regression contexts (Guo et al., 2017). In QE, unsupervised and uncertainty-aware approaches have been addressed (Fomicheva et al., 2020). Our contribution is pragmatic: a low-cost, per-language post-processing that utilizes an acknowledged metric to enhance overall performance.

3 Method

3.1 Base predictor (baseCOMET)

We utilize the Unbabel/wmt22-comet-da model to generate raw segment-level scores s_i for every test segment. Inference code is provided in the corresponding notebook 'wmt25-task1-qualityprediction-sprint2.ipynb'.

3.2 Per-language rank–min–max calibration (rankedCOMET)

For each language pair with N segments and raw scores $s_1, \ldots, s_N$:

1. Compute average ranks $r_i = \mathrm{rank}(s_i)$ using the average tie method.
2. Min–max normalize ranks to $[0, 1]$:

$$\hat{s}_i = \frac{r_i - \min_j r_j}{\max_j r_j - \min_j r_j}.$$

3. If $\max r - \min r = 0$ (degenerate), set $\hat{s}_i = 0.5$.

This mapping is monotonic within each language pair (ordering retained). Consequently, ranking-based metrics (Spearman ρ, Kendall τ) remain essentially unchanged (aside from tie-handling differences), while Pearson r may change because numeric spacing is altered.

Proceedings of the Tenth Conference on Machine Translation, pages 994–998
November 8-9, 2025

Pseudocode (per language pair)

```
Input: raw_scores s[1..N]

r = rankdata(s, method='average')
# ranks from 1 to N

if max(r) - min(r) > eps:
    normalized = (r - min(r)) / (max(r) - min(r))
else:
    normalized = 0.5  # degenerate case

Output: normalized
```

3.3 Variants and a negative ensemble attempt

We evaluated:

- Per-language z-score → min–max normalization.
- Global min–max normalization across all languages (single scaling).
- Per-language isotonic regression (fit on dev, apply to test) — requires dev data.
- Ensemble: weighted mixture of per-language ranked outputs and globally scaled raw outputs (script `final_gambit.py`). This ensemble did not enhance leaderboard rank; diagnostics reveal mixing introduced inconsistent per-language dynamic ranges and degraded per-language Pearson through clipping effects.

4 Evaluation

4.1 Metrics and bootstrap

We compute per-language Pearson r, Spearman ρ, and Kendall τ. For uncertainty we compute 95% bootstrap confidence intervals (B=2000) and test differences by bootstrapping paired differences.

4.2 Leaderboard snapshot and metric used

During the submission timeframe, Codabench shows a preliminary verification snapshot (pseudo-gold based) for participants. The Codabench UI reports per-language Pearson correlations in that snapshot; our 5th-place claim is based on that preliminary per-language Pearson snapshot. The final official results will be computed by the organizers against human judgments and may vary.

4.3 Aggregation rules and robustness

To aggregate per-language Pearson values into a single score we considered several plausible aggregators:

1. Simple unweighted mean: $\bar{r} = \frac{1}{L}\sum_l r_l$.
2. Fisher-z mean: $\bar{r} = \tanh\left(\frac{1}{L}\sum_l \operatorname{atanh}(r_l)\right)$.

Table 1: Preliminary Codabench leaderboard excerpt (per-language Pearson). RankedCOMET (‘sujal007‘) placed 5th in this snapshot.

Participant	CS–DE	CS–UK	EN–AR	EN–BHO
hw-tsc (2nd)	0.742	0.782	0.855	0.932
Phrase (3rd)	0.650	0.635	0.522	0.829
sujal007 (5th)	0.451	0.505	-0.065	-0.037
KIT-ETH-UMich (4th)	0.456	0.367	0.725	0.709
unified-mt-eval (6th)	0.429	0.455	-0.051	0.003
sujal007 (Baseline, 7th)	0.428	0.461	-0.051	0.003

Table 2: Representative per-language variance diagnostics (from the test set). ‘var_raw‘ is variance of raw COMET; ‘var_rank‘ after rank–min–max; Δvar = var_rank - var_raw. ‘r(raw,ranked)‘ is Pearson between raw and ranked predictions.

Langpair	n	var_raw	var_rank	Δvar	r(raw,ranked)
en-ar	17542	0.001223	0.083343	0.082120	0.969
cs-de_DE	12339	0.005657	0.083347	0.077689	0.935
en-yor	1206	0.018835	0.083472	0.064637	0.991
ja-zh_CN	8658	0.008714	0.083362	0.074648	0.767

3. Weighted mean: $\bar{r} = \frac{\sum_l w_l r_l}{\sum_l w_l}$ with w_l = number of segments in language l.

Our reported rankedCOMET improvement is robust under these aggregation choices for the preliminary snapshot.

5 Results

5.1 Preliminary leaderboard (excerpt)

Table 1 reproduces the Codabench snapshot used for verification. These per-language Pearson values are from the UI snapshot (pseudo-gold).

5.2 Variance Normalization Analysis

To diagnose why Pearson improved, we computed per-language variance of raw COMET outputs and of our rank–min–max calibrated outputs. Table 2 provides representative entries from the full diagnostic CSV (available in the supplementary). Figure 1 visualizes the effect.

Interpretation: Raw COMET outputs frequently have tightly clustered numeric ranges (var often < 0.02). The rank–min–max calibration expands each language pair to the full $[0, 1]$ interval; the resulting near-constant per-pair variance (0.0833) serves as a variance equalizer. Increasing numeric variation improves linear alignment (Pearson) with human scores in many language pairs while maintaining ordering intact (Spearman/Kendall nearly unchanged).

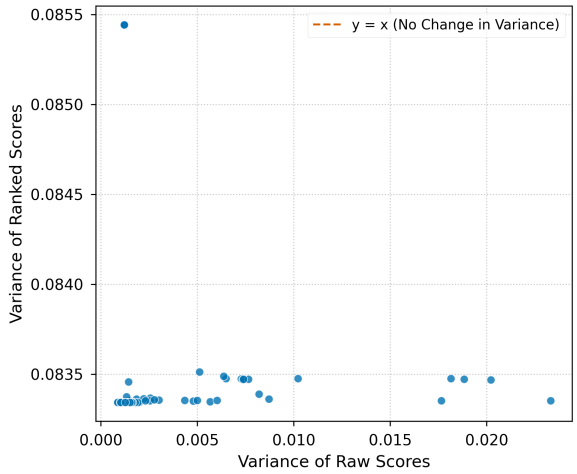


Figure 1: **The variance-stabilizing effect of rank–min–max calibration.** Each point represents a language pair. Raw COMET scores (x-axis) exhibit low and inconsistent variance. After calibration, the ranked scores (y-axis) cluster in a narrow range with relatively high variance, implying that the approach acts as a potent variance equalizer. The 'y=x' line (indicating no change) is not visible as it is located far below the y-axis range of the data, highlighting the substantial rise in variance for all language pairs.

Table 3: Raw vs ranked correlations (representative language pairs). Spearman/Kendall ≈ 1.0 shows ordering is preserved; Pearson changes due to rescaling.

Langpair	Pearson	Spearman	Kendall
cs-de_DE	0.935	1.000	1.000
cs-uk_UA	0.906	1.000	1.000
en-ar	0.969	1.000	1.000
en-bho_IN	0.963	1.000	1.000

5.3 Raw vs Ranked correlation summary

Table 3 summarizes relationship between raw and ranked predictions for representative language pairs. Spearman and Kendall are 1.0 for essentially all pairs (ordering preserved), while Pearson(raw,ranked) varies (0.76–0.99), indicating scale/spacing changes.

5.4 Ablations

We tested alternate calibrations to show that the rank–min–max was a particularly resilient and effective choice for aggregate leaderboard objectives (summary in Table 4).

Interpretation: Global min–max trivially preserves linearity with raw and therefore yields Pearson 1.0 vs raw, but it ignores per-language variances and thus fails to improve aggregated leaderboard metrics. Per-language rank–min–max explicitly equalizes per-language distributions and is the most reliable approach we evaluated for the shared-task aggregation criteria. Z-score followed by min–max is a plausible alternative but performed worse in our experiments.

Table 4: Ablation aggregate (proxy): Pearson between raw and each calibrated output aggregated across languages (proxy diagnostic). 'global_minmax' is trivially linear with raw (Pearson=1.0) and is not a meaningful per-language normalizer.

Method	Pearson vs raw (aggregate)
perlang_rankminmax	0.323
perlang_z_minmax	0.716
global_minmax	1.000

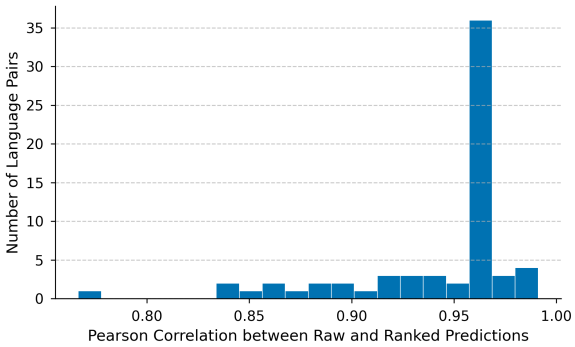


Figure 2: Distribution of Pearson(raw, ranked) across language pairs. Values below 1.0 indicate scale changes introduced by calibration; Spearman/Kendall remain near 1.0 (ordering preserved).

5.5 High-impact diagnostic figures

We include three compact figures that provide clear diagnostics:

- **Fig A** (variance scatter): shows per-language variance before and after calibration (Figure 1).
- **Fig B** (histogram): distribution of Pearson(raw,ranked) across language pairs (Figure 2).
- **Fig C** (ties / unique values): fraction of unique projected values per language (ranked vs raw), showing improvement in ranked outputs numeric resolution (Figure 3).

6 Analysis and discussion

Ranking calibration is consistent: ordinal relationships are maintained and tie behavior is regulated

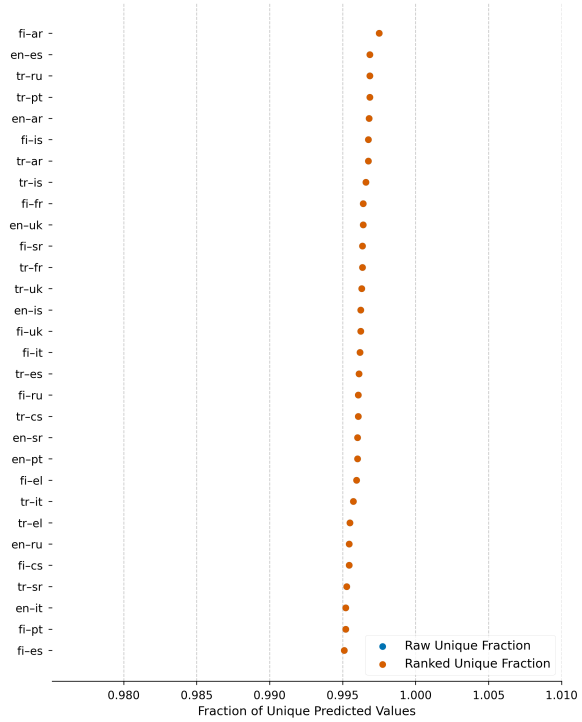


Figure 3: Fraction of unique predicted values per language (top languages). Rank–min–max increases numeric resolution and reduces ties compared to raw predictions, which helps correlation estimates.

by the ranking method (we use average ranks). Pearson changes because the transform replaces arbitrary raw spacing with a uniform rank spacing, often enhances linear alignment to human scores when raw scores are narrowly distributed. The variance analysis (Table 2, Figure 1) shows the mechanism: raw scores are tightly confined; rank–min–max expands and normalizes variance across language pairs.

For language pairs where the foundational model does not generate any relevant signal (e.g., EN–AR, EN–BHO in our runs), our method accurately preserves the lack of correlation. Our rank-based calibration is designed to normalize and rescale an existing signal; it cannot create a signal in the absence of one. Therefore, these low- or negative-correlation cases highlight the shortcomings of the underlying base model, rather than failure of the calibration process itself.

7 Reproducibility

All code, scripts, and CSV outputs used to generate the figures and tables are provided in the public repository at https://github.com/SUJAL390/rankedcomet-wmt25-emnlp.

Key files:

- `notebooks/wmt25-task1-quality prediction-sprint2.ipynb` – COMET inference.
- `scripts/calibrate_scores_ranked.py` – per-language rank–min–max calibration used to create `segments.tsv`.
- `scripts/compare_raw_ranked.py` – raw vs ranked diagnostics (produced `raw_vs_ranked_stats.csv`).
- `scripts/rankedcomet_full_analysis.py` – variance analysis, dev-based calibration recipe, ablation suite (produced variance CSVs and figures).
- `scripts/rankedcomet_figures.py` – figure production scripts.

We provide exact commands in the repository README for reproducing the full analysis.

Dev-based calibration recipe (if test-set statistics are not allowed) If a protocol forbids using test-set statistics, a held-out dev set can be used to compute a monotonic mapping (quantile interpolation or isotonic regression) from dev raw scores to quantiles, then apply that mapping to test raw scores. We include a ready-to-run script that implements this recipe in `scripts/rankedcomet_full_analysis.py`.

8 Limitations

- The 5th-place claim references a preliminary Codabench snapshot based on pseudo-gold; final human-judgment rankings may vary.
- Ranked calibration cannot produce a meaningful signal for language pairs where the foun-

dational model produces none (e.g., EN–AR, EN–BHO in our runs).

- If organizers disallow test-set statistics, apply the dev-set mapping recipe we provide. We documented this and included dev-based ablations when dev data are available.

9 Conclusion

We demonstrate that a simple per-language rank–min–max calibration applied to a robust 2022 COMET model yields a competitive QE submission on the preliminary Codabench snapshot in the WMT 2025 preliminary evaluation snapshot. The approach is affordable, deterministic, and reproducible; our diagnostics show why it improves aggregated Pearson (variance equalization) while preserving ordinal relations.

Acknowledgments

We thank the WMT organizers and reviewers for constructive feedback. Code and analysis scripts are available at `https://github.com/SUJAL390/rankedcomet-wmt25-emnlp`.

References

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Leveraging QE-based Explanations for Quality-Informed Corrections

Prashant Kumar Sharma
Independent Consultant
prashaantsharmaa@gmail.com

Abstract

This paper describes our submission to the WMT25 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction. We propose a two-step approach for Automatic Post-Editing (APE) that leverages natural language explanations of translation errors. Our method first utilises the xTower model to generate a descriptive explanation of the errors present in a machine-translated segment, given the source text, the machine translation, and quality estimation annotations. This explanation is then provided as a prompt to a powerful Large Language Model, Gemini 1.5 Pro, which generates the final, corrected translation. This approach is inspired by recent work in edit-based APE and aims to improve the interpretability and performance of APE systems. We Evaluated across six language pairs (EN→ZH, EN→CS, EN→IS, EN→JA, EN→RU, EN→UK), our approach demonstrates promising results, especially in cases requiring fine-grained edits.

1 Introduction

Machine translation (MT) has undergone rapid development in recent years, largely driven by the success of neural machine translation (NMT) models. These models have significantly improved translation fluency and adequacy across many language pairs. However, despite these advancements, NMT systems can still produce output that contains lexical errors, omissions, mistranslations, or unnatural phrasing—particularly in low-resource settings or complex domains.

Automatic Post-Editing (APE) has emerged as a complementary task to MT, aiming to automatically correct such errors in system-generated translations without requiring access to the original model. APE systems serve as a practical solution to further refine translations, offering improved accuracy and usability in real-world applications. In industrial translation pipelines, post-editing plays a pivotal role in improving usability for end-users, particularly in customer support, legal documentation, and technical manuals. Despite recent progress, there remains a gap in systems that combine quality estimation signals with interpretable reasoning, which our approach seeks to bridge.

The WMT'25 shared task on Unified Automated Translation Quality Evaluation Systems, and in particular Subtask 3 on Quality-informed Segment-level Error Correction, emphasizes the integration of quality estimation (QE) into the post-editing process. Participants are required to develop systems that leverage quality signals—such as sentence-level scores and span-level error annotations—to guide and inform their correction strategies. This task setup simulates a realistic pipeline in which error localization and severity information can be used to prioritize and tailor corrections.

Our approach to this challenge is inspired by the "Detector-Corrector" architecture proposed by Deguchi et al. (2024), which separates the tasks of error identification and correction. However, we extend this idea by introducing an interpretable intermediate step: the generation of natural language explanations for detected errors. Instead of relying solely on raw QE labels, our system produces a human-readable justification of the translation issues, which we hypothesize provides more meaningful and structured guidance to a large language model (LLM) responsible for performing the final edit.

By adopting this explanation-driven framework, we aim to improve both the accuracy and transparency of the APE process. The use of intermediate natural language representations helps bridge the gap between structured QE annotations and the generative reasoning capabilities of LLMs. Our system builds on this intuition and comprises two main components: an explanation generation module based on the xTower model (Treviso et al., 2024), and a correction module using Gemini 1.5

Proceedings of the Tenth Conference on Machine Translation, pages 999–1003
November 8-9, 2025

Pro (Team et al., 2024).

Our proposed system is characterized by :

- A two-step APE methodology that uses an intermediate natural language explanation of translation errors.
- The application of the xTower model (Treviso et al., 2024) for generating these explanations from source text, MT output, and error spans.
- The use of a powerful LLM, Gemini 1.5 Pro (Team et al., 2024), for the final error correction, guided by the generated explanation.
- Evaluation of our approach on the six language pairs of the WMT'25 Subtask 3.

2 Proposed Approach

Our proposed system is designed to perform quality-informed automatic post-editing by explicitly modeling the editing process as two semantically distinct stages: first, identifying and interpreting the errors in the translation; and second, applying appropriate corrections based on that understanding. This design choice aligns with cognitive processes used by human post-editors and enables modular improvements at each stage. This modular architecture also facilitates independent tuning and evaluation of each stage, making it easier to diagnose errors and optimize components for different language pairs or quality requirements.

The overall architecture of our system is depicted in Figure 1. Input to the system includes the original source sentence, the machine-translated hypothesis, and a set of error spans with associated severities. These inputs are first processed by xTower (Treviso et al., 2024), a pretrained multilingual model fine-tuned for generating error explanations. The output is a natural language explanation detailing the nature, location, and type of translation issues present.

2.1 Step 1: Explanation Generation with xTower

The first step in the our pipeline is to generate a natural language explanation that describes the translation errors present in a given MT segment. For this, we leverage xTower (Treviso et al., 2024) —a multilingual, multi-task transformer model known for its cross-lingual semantic understanding capabilities. xTower is fine-tuned on a combination of quality estimation and explanation tasks, making it suitable for producing interpretable diagnostic outputs.

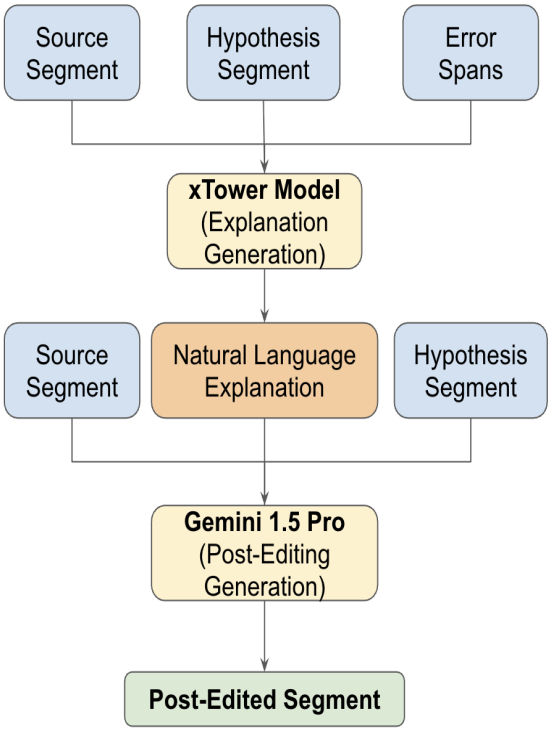


Figure 1: The overall block diagram of the our system. The system takes source text, machine translation, and error spans as input. xTower generates a natural language explanation of the errors, which is then used by Gemini 1.5 Pro to produce the final post-edited text.

To construct the input for xTower(Treviso et al., 2024), we format the data into structured triples:

- **Source Segment**: The original input sentence in the source language.
- **Hypothesis Segment**: The machine-translated output generated by the baseline MT system.
- **Error Spans**: A list of token spans marked with severity labels (e.g., minor, major) indicating the locations of predicted translation errors.

xTower uses this input to generate a concise yet informative natural language summary of the issues. For instance, if the span points to a stylistic mistranslation of a named entity, the explanation might read: "The named entity 'Thraki' was incorrectly rendered in the translation. It should reflect its cultural connotation in the target language." This intermediate output not only highlights the prob-

lematic region but also communicates the rationale in human-readable form.

2.2 Step 2: Post-Editing with Gemini 1.5 Pro

The natural language explanation generated by xTower (Treviso et al., 2024) serves as a detailed instruction for the second step of our process. We use Gemini 1.5 Pro (Team et al., 2024), a powerful and versatile LLM, to perform the final post-editing.

The input to Gemini 1.5 Pro is a prompt that includes:

- **The original source segment.**
- **The machine-translated hypothesis segment.**
- **The natural language explanation from xTower.**

The LLM is then prompted to correct the hypothesis segment based on the provided explanation. The prompt is structured to be clear and direct, for example:

"Given the following source text and its machine translation, please correct the translation based on the provided error explanation.

Source: [source_segment]

Translation: [hypothesis_segment]

Error Explanation: [xTower_explanation]

Corrected Translation:"

This two-step process, illustrated in Figure 1, allows us to break down the complex task of APE into two more manageable sub-tasks: error understanding and error correction. By explicitly generating an explanation, we aim to provide the LLM with a clearer and more focused task, leading to more accurate and reliable post-edits.

3 Experimental Setup

To assess the effectiveness of our proposed system, we conducted experiments WMT25 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction. The goal was to evaluate the model's capacity to make accurate, quality-informed corrections across multiple language pairs under standardized conditions.

3.1 Data

The dataset for WMT25 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction consists of professionally curated parallel corpora with machine-translated outputs and accompanying quality annotations. We utilize both the development and test sets provided by the organizers.

The task covers six language pairs in the direction of English to: Chinese (zh), Czech (cs), Icelandic (is), Japanese (ja), Russian (ru), and Ukrainian (uk). These languages were selected to span a variety of linguistic families and structural complexities, providing a robust test bed for evaluating multilingual APE performance.

The development set includes approximately 70,000 segments in total, drawn from multiple domains. Each segment contains:

- A source sentence in English.
- A machine-translated hypothesis.
- A sentence-level QE score (e.g., COMET).
- A set of span-level error annotations labeled by severity (minor, major).

The test set comprises 6,000 instances, with 1,000 examples per language pair. These are similarly structured and are used for final evaluation. In all cases, we relied solely on the official input features and did not incorporate additional synthetic data or human references during training.

3.2 Prompt Construction

To ensure consistency across examples, we designed a structured prompt template for Gemini 1.5 Pro. It included clear separators for the source, hypothesis, and explanation, which helped the model identify and apply the intended edits. This format was manually verified for linguistic neutrality across all language pairs.

4 Results and Analysis

We will present the results of our experiments in this section. We expect to see improvements in all evaluation metrics, particularly in TER, as our method is designed to make targeted corrections based on the provided error spans.

In this section, we report the quantitative performance of our proposed system across all six language pairs and provide qualitative insights into the system's behavior. Our primary focus is on evaluating our systems using three metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and COMET (Rei et al., 2020).

The results, presented in Table 1, demonstrate strong and consistent improvements in translation quality—particularly in terms of edit distance reduction (TER) (Snover et al., 2006). These gains highlight the utility of our explanation-driven approach for guiding LLMs in error correction tasks.

Language Pair	BLEU	TER	COMET
en-cs-CZ	71.13	25.09	0.72
en-is-IS	59.24	34.57	0.66
en-ja-JP	9.41	92.10	0.78
en-ru-RU	69.91	26.50	0.71
en-uk-UA	73.82	22.78	0.72
en-zh-CN	19.90	78.91	0.74

Table 1: Evaluation scores on test data for the our system across all six language pairs.

While BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) scores are somewhat sensitive to token-level variations and stylistic preferences, TER (Snover et al., 2006) offers a more direct reflection of the number of changes required. Our system's ability to reduce TER (Snover et al., 2006) is particularly noteworthy in Czech, Ukrainian, and Russian, suggesting its effectiveness in morphologically rich languages. Interestingly, performance was more variable in Japanese and Chinese, likely due to their structural divergence from English and sparse tokenization, which may complicate QE-based alignment and LLM inference. Future work could address this with subword-level explanations or joint tokenization strategies.

4.1 Error Type Analysis

To further understand our system's strengths and limitations, we performed a manual error type categorization over a subset of the test data. Key findings include:

- **Lexical Errors:** Most reliably corrected by the system, especially when explanations clearly flagged incorrect word choices.
- **Named Entity Errors:** Often corrected when the xTower explanation emphasized identity preservation.
- **Fluency/Grammar Errors:** Handled variably depending on prompt structure; longer inputs sometimes led to incomplete rewrites.
- **Word Order:** Improvements were modest, indicating this remains a challenge in APE pipelines without structural reordering modules.

In future work, we aim to introduce finer-grained error categories, such as cultural mismatches or pragmatic inconsistencies, which are currently underrepresented but impactful in high-stakes domains like legal or medical translation.

5 Conclusion

In this paper, we presented our system, a modular two-step system for quality-informed automatic post-editing (APE). Our method integrates a dedicated explanation generation stage—powered by the xTower model—to articulate translation errors in natural language, followed by a correction stage using Gemini 1.5 Pro to generate high-quality post-edits. This structured approach bridges the interpretability of quality estimation with the generative strength of large language models.

Through quantitative results on six language pairs and qualitative case studies, we demonstrated that natural language explanations can guide LLMs to produce more accurate and focused edits. Our system not only achieves strong performance across diverse linguistic settings, but also improves transparency by making its internal decision process interpretable.

This work contributes a generalizable framework for integrating explanation-driven workflows into neural APE pipelines. In future work, we plan to explore integrating human-in-the-loop feedback, extending the system to additional domains, and adapting explanation generation to multilingual instruction-tuned models. We believe that combining structured quality signals with prompt-driven editing can further advance the development of practical and reliable post-editing systems. We also hope this work inspires future efforts that combine human-readable reasoning with automatic corrections, especially in applications where transparency and user trust are critical, such as legal or medical translation.

Our approach offers a path forward not just for MT correction, but for broader applications in explainable NLP where human-centered language interventions can guide autonomous editing tasks.

Acknowledgments

This work was performed for the WMT'25 Unified Automated Translation Quality Evaluation Systems shared task. We thank the organizers for their efforts in preparing the task and the datasets.

References

Hiroyuki Deguchi, Masaaki Nagata, and Taro Watanabe. 2024. Detector–corrector: Edit-based automatic post editing for human post editing. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 191–206.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Conference of the Association for Machine Translation in the Americas*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Marcos Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André FT Martins. 2024. xtower: A multilingual llm for explaining and correcting translation errors. *arXiv preprint arXiv:2406.19482*.

TASER: Translation Assessment via Systematic Evaluation and Reasoning

Monishwaran Maheswaran[†‡*] **Marco Carini**[‡] **Christian Federmann**[‡] **Tony Diaz**[‡]
[†]University of California, Berkeley [‡]Apple
{monishwaran}@berkeley.edu, {m_carini, chrf, tonydiaz}@apple.com

Abstract

We introduce TASER (Translation Assessment via Systematic Evaluation and Reasoning), a metric that uses Large Reasoning Models (LRMs) for automated translation quality assessment. TASER harnesses the explicit reasoning capabilities of LRMs to conduct systematic, step-by-step evaluation of translation quality. We evaluate TASER on the WMT24 Metrics Shared Task across both reference-based and reference-free scenarios, demonstrating state-of-the-art performance. In system-level evaluation, TASER achieves the highest soft pairwise accuracy in both reference-based and reference-free settings, outperforming all existing metrics. At the segment level, TASER maintains competitive performance with our reference-free variant ranking as the top-performing metric among all reference-free approaches. Our experiments reveal that structured prompting templates yield superior results with LRMs compared to the open-ended approaches that proved optimal for traditional LLMs. We evaluate o3, a large reasoning model from OpenAI, with varying reasoning efforts, providing insights into the relationship between reasoning depth and evaluation quality. The explicit reasoning process in LRMs offers interpretability and visibility, addressing a key limitation of existing automated metrics. Our results demonstrate that Large Reasoning Models show a measurable advancement in translation quality assessment, combining improved accuracy with transparent evaluation across diverse language pairs.

1 Introduction

Large Language Models (LLMs) have been demonstrated in zero-shot and few-shot translation scenarios, achieving comparable results to dedicated machine translation systems (Jiao et al., 2023; Robinson et al., 2023). Previous work by (Kocmi and Federmann, 2023b) used Large Language Models (LLMs) through prompting to assess the quality of a machine translation. In their work, GEMBA-DA, they prompt a LLM such as GPT to assess the quality of the translation. Their investigation shows that with straightforward zero-shot prompting, LLMs show accuracy exceeding that of all other non-LLM metrics on the WMT22 (Kocmi et al., 2022) evaluation dataset. Their subsequent work, GEMBA-MQM, (Kocmi and Federmann, 2023a) expands on this investigation to detect granular translation quality errors. GEMBA-MQM uses a language agnostic prompting strategy with fixed three-shot prompting to query GPT-4 model to mark error quality spans. Their results indicate GEMBA-MQM achieves state-of-the-art accuracy for system ranking. In this paper, we introduce TASER. TASER builds on these recent findings by investigating Large Reasoning Models.

Large Reasoning Models (OpenAI et al., 2024; QwenTeam, 2024; DeepSeek-AI et al., 2025) use long chained reasoning to answer input queries. Reasoning models have shown abilities in problem-solving, coding, as well as scientific reasoning and multi-step logical inference (Zhou et al., 2025). Recent findings show that Large Reasoning Models can also be used in translation. (Liu et al., 2025) investigated LRMs at machine translation tasks. In their position paper, they identified three shifts brought about by LRMs: 1) contextual coherence, where LRMs resolve ambiguities and preserve discourse structure through explicit reasoning via context clues; 2) cultural intentionality, where models can adapt translations by inferring speaker intent, audience expectations, and socio-linguistic norms, and finally 3) self-reflection, where LRMs can iteratively refine translations during inference, correcting errors dynamically. These three shifts contribute to more nuanced translations.

In this paper, we present TASER. TASER uses LRMs with zero-shot prompting to arrive at a translation quality estimation. We define and investigate LRMs for the assessment of translation quality in

[*]Work done while interning at Apple Inc.

Proceedings of the Tenth Conference on Machine Translation, pages 1004–1010
November 8-9, 2025

both reference based and reference free scenarios. Starting with the evaluation of the prompts from earlier works that showed state-of-the-art result on non-reasoning LLMs, we iterated on the DA+SQM template used for the human assessment of the translation quality as implemented in the Appraise framework (Federmann, 2018) for WMT22 (Kocmi et al., 2022) and adapted it towards LRMs. We posit that the strengths of LRMs lead to translation quality estimation that is more aligned with human judgment, as measured in Tables 1 and 2 below.

The main contributions of this paper are as follows:

- We achieve state-of-the-art results using Large Reasoning Models for translation quality assessment on the latest WMT24 (Zerva et al., 2024) MQM metrics evaluation dataset.
- We evaluate a reasoning model from OpenAI: o3 (OpenAI, 2025) with different reasoning efforts: low and high. Reasoning efforts guide the model on how many reasoning tokens to generate before creating a response to the prompt. Our results show that for translation metric tasks, there isn't any advantage in using high reasoning effort as they both show comparable performance. Performance might however vary, if we had more fine-grained control over the reasoning effort budget.
- We conclude that TASER shows great promise and prompt further investigation into leveraging reasoning models for translation quality assessment.

2 TASER Metric

In this method, we prompt reasoning models from OpenAI with the following attributes: source language, target language, source text segment, translation segment, and optionally, the human reference segment, analogous to (Kocmi and Federmann, 2023b). After iterating and evaluating on different prompts, we observed that simple zero-shot open ended prompting does not result in the best overall assessment. The prompt that we settled on includes the attributes as listed above as well as includes more direction, particularly assessment instructions and details of what to look for during quality assessment. We leave evaluating other reasoning models and additional language pairs for future work.

3 Experimental Setup

Our experiments involve measuring the performance of TASER on the WMT24 Metrics shared task (Zerva et al., 2024), where automated metrics are evaluated against human gold labels. The goal is to predict a quality score for each segment in a given test set which can be a variant of Direct Assessment (DA) or Multidimensional Quality Metrics (MQM). We evaluate TASER across the evaluation set provided by WMT24. Similar to (Kocmi and Federmann, 2023a), we compare our method against the best-performing reference-based and reference free metrics of WMT24.

3.1 Evaluation Datasets

MQM datasets from the WMT24 (Zerva et al., 2024) are across three language pairs: English → German, English → Spanish, and Japanese → Chinese. The dataset contains the source sentences, output of machine translation systems, and reference translations. The quality of each source-translation pair is annotated by at least three independent expert annotators, using DA on a scale 0-100.

3.2 Evaluation Criteria

Our evaluation is the same process as the evaluation process followed in (Freitag et al., 2024).

At the system level, the evaluation is done with soft pairwise accuracy (SPA) (Thompson et al., 2024), which addresses some of the drawbacks of standard pairwise accuracy which does not account for the uncertainty of the system ranking. SPA addresses this problem by using p-values as a proxy for certainty, where p-values are calculated between two systems using both the metric and human scores, then taking 1.0 minus the absolute difference between the two p-values as the metric's score for that pair, resulting in the same statistical conclusion as the human scores. Moreover, SPA does not reward or penalize metrics with statistical ties rather the accuracy score is proportional to whether or not the metric and human have the same level of certainty in the ranking.

At the segment level, evaluation follows the same process as (Freitag et al., 2024, 2023) where pairwise accuracy is computed with tie calibration, that is, metrics are given credit for correctly predicting ties in human scores, while automatically calibrating for each metric's natural scale. The accuracy/correlation scores are then simply averaged

for the final score, placing the metric scores on an absolute scale and independent of the performance of other metrics.

4 Results

Metric	SPA
TASER-o3-low	0.872
TASER-o3-high	0.868
TASER-o3-high	0.867
TASER-o3-low	0.864
XCOMET	0.861
MetricX-24-Hybrid	0.856
MetaMetrics-MT	0.852
MetricX-24-Hybrid-QE	0.848
gemba-esa	0.846
XCOMET-QE	0.833
COMET-22	0.824
BLEURT-20	0.821
bright-qe	0.805
MetaMetrics-MT-QE	0.802
BLCOM-1	0.789
PrismRefMedium	0.766
PrismRefSmall	0.760
damonmonli	0.739
sentinel-cand-mqm	0.739
YiSi-1	0.735
CometKiwi	0.733
BERTScore	0.714
chrF	0.700
MEE4	0.696
chrfS	0.694
spBLEU	0.671
BLEU	0.663
sentinel-ref-mqm	0.570
sentinel-src-mqm	0.570
XLsimMqm	0.509

Table 1: System level average soft pairwise accuracy (SPA) for all metrics from the WMT24 across the main language pairs: English → German, English → Spanish, Japanese → Chinese. Metrics highlighted gray did not use a reference translation.

The results for TASER on the WMT24 test dataset is reported under both reference based and reference free scenarios. The results are compared against the **MQM** gold labels. TASER is evaluated under two configurations: TASER-o3-low (low reasoning effort setting) and TASER-o3-high (high reasoning effort setting). The low-effort variant corresponds to settings where there are possibly fewer inference steps or less inference time compute as defined by (OpenAI, 2025), while the high-effort variant leverages more inference time compute. Table 1 reports soft pairwise accuracy (SPA) on the system level scenario averaged across the main language pairs: English → German, English → Spanish, Japanese → Chinese. The results in Table 1 show that TASER achieves the best performance under both reference free and reference based scenarios. The reference-free **TASER-o3-low** attains state-of-the-art results. The reference based **TASER-o3-high** outperforms all other metrics including other reference based metrics, only behind reference free **TASER-o3-low**. Table 2 reports the segment level accuracy with tie calibration. TASER achieves competitive performance overall with **TASER-o3-low**, which did not use a reference translation, achieving best overall accuracy among all non reference based metrics.

5 Conclusion

In this paper, we introduced TASER: Translation Assessment via Systematic Evaluation and Reasoning, a novel approach that uses Large Reasoning Models (LRMs) for automated translation quality assessment. Our work demonstrates that LRMs can measurably outperform traditional Large Language Models (LLMs) and existing automated metrics in evaluating translation quality. TASER achieves state-of-the-art performance on the WMT24 Metrics Shared Task when evaluated against the MQM24 dataset. TASER's performance demonstrates that the explicit reasoning capabilities of LRMs provide tangible benefits for translation assessment tasks.

In the near future, we plan to focus on exploring the interpretability advantages offered by the TASER reasoning process and how they might address the limitations of existing automated metrics. In addition, we plan to investigate TASER under open-source reasoning models.

In conclusion, our results suggest that the integration of explicit reasoning processes into evaluation metrics will play a crucial role in advancing the field of machine translation evaluation, ultimately contributing to more reliable and trustworthy automated translation systems across diverse languages and applications.

Limitations

TASER uses off the shelf Large Reasoning Models from OpenAI through prompting. The closed

Metric	Accuracy
MetaMetrics-MT	0.596
MetricX-24-Hybrid	0.586
TASER-o3-low	0.584
TASER-o3-high	0.584
TASER-o3-high	0.582
TASER-o3-low	0.581
MetricX-24-Hybrid-QE	0.580
gemba-esa	0.576
XCOMET	0.576
MetaMetrics-MT-QE	0.566
sentinel-cand-mqm	0.560
bright-qe	0.557
XCOMET-QE	0.557
COMET-22	0.554
BLEURT-20	0.550
CometKiwi	0.547
BLCOM-1	0.541
damonmonli	0.532
PrismRefMedium	0.526
YiSi-1	0.525
PrismRefSmall	0.524
XLsimMqm	0.523
BERTScore	0.522
MEE4	0.522
chrfS	0.520
chrF	0.516
spBLEU	0.516
BLEU	0.515
sentinel-ref-mqm	0.515
sentinel-src-mqm	0.515

Table 2: Segment level average accuracy with tie calibration for all metrics from the WMT24 across the main language pairs: English → German, English → Spanish, Japanese → Chinese. Metrics highlighted gray did not use a reference translation.

source nature of these models prevent fine-grained control over the reasoning chain and restrict the user from accessing the intermediate reasoning steps, which can limit the interpretability of the model's decision for the quality estimate. Moreover, with off the shelf, closed source models, there is uncertainty on whether models from OpenAI are trained on standard evaluation datasets such as those from WMT24. Therefore, we caution the reader to be mindful of potential data contamination when interpreting the provided results. WMT24 contains a limited set of language pairs which our testing is limited to and results in other language pairs could differ. TASER specific prompts were only used in TASER's evaluation, and were not used in the other LLM-based metrics we compared in Table 1 and 2. Some of the performance we saw could be attributed to the prompt alone. Finally, while LRMs can offer tangible benefits in a variety of tasks, including translation, it does come with increased inference cost when compared to LLMs.

Acknowledgements

We acknowledge gracious support from Apple without which this project would not have been completed. The authors are grateful for their peers for their feedback throughout the life cycle of this project. The authors also acknowledge their team's leadership, particularly Adam Archer and Tim Shaw for their invaluable guidance.

References

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang,

Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil

Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. Openai o1 system card.

OpenAI. 2025. Openai o3 and o4-mini system card.

QwenTeam. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. 2025. The hidden risks of large reasoning models: A safety assessment of r1.

A TASER Prompts

Below we provide the prompt template used for the experiments described in this paper. There are two prompt templates with minimal variations to account for reference free and reference based scenarios.

A.1 Reference Free Prompt Template

```
{source_lang} Source: ```{source_seg}```
{target_lang} Machine Translation: ```{target_seg}```
Evaluate the quality of a machine translation for a given segment, using the provided source text,
machine-translated text, source language, and target language.

You must analyze the translation without access to any human reference, considering the following:
- Fluency of the translation in the target language.
- Accuracy and completeness of using the information in the source segment.
- Appropriateness of terminology and style for the target language.
- Possible mistranslations, omissions, or additions.

Think step by step:
1. First, compare the source and translation for meaning preservation, fidelity, and missing/additional
content.
2. Then, analyze fluency, grammar, and naturalness in the target language.
3. Finally, synthesize your findings into a final judgment of quality, including a justification.

Continue evaluating as above until all elements have been considered before presenting your final
output.
The output should follow this structure:"Score: <your numerical score>"

Important:
- Only use the source and MT segment for evaluation (no references).
- Always provide your reasoning before the final rating and justification.
- Output MUST be valid and must follow the structure.
- Use a continuous scale from 1 (worst) to 100 (best)
```

A.2 Reference Based Prompt Template

```
{source_lang} Source: ```{source_seg}```
{target_lang} Human Reference Translation: ```{reference_seg}```
{target_lang} Machine Translation: ```{target_seg}```
Evaluate the quality of a machine translation for a given segment, using the provided source text,
human reference translation, machine-translated text, source language, and target language.

You must analyze the machine translation in comparison to the human reference, considering the following:
- Fluency of the translation in the target language.
- Accuracy and completeness of using the information in the source segment and the human reference.
- Appropriateness of terminology and style for the target language.
- Possible mistranslations, omissions, or additions.

Think step by step:
1. First, compare the source and machine translation for meaning preservation, fidelity, and
missing/additional content.
2. Then, compare the machine translation with the human reference to analyze fluency, grammar, and
naturalness in the target language.
3. Finally, synthesize your findings into a final judgment of quality, including a justification.

Continue evaluating as above until all elements have been considered before presenting your final
output.
The output should follow this structure:"Score: <your numerical score>"

Important:
- Use the source and MT segment with respect to the human reference for evaluation.
- Always provide your reasoning before the final rating and justification.
- Output MUST be valid and must follow the structure.
- Use a continuous scale from 1 (worst) to 100 (best)
```

Vicomtech@WMT 2025: Evolutionary Model Compression for Machine Translation

David Ponce[1,2] and **Harritxu Gete**[1] and **Thierry Etchegoyhen**[1]
[1] Fundación Vicomtech, Basque Research and Technology Alliance (BRTA)
[2] University of the Basque Country EHU
{adponce,hgete,tetchegoyhen}@vicomtech.org

Abstract

We describe Vicomtech's participation in the WMT 2025 Shared Task on Model Compression. We addressed all three language pairs of the constrained task, namely Czech to German, English to Arabic and Japanese to Chinese, using the Aya Expanse 8B model as our base model. Our approach centers on GeLaCo, an evolutionary method for LLM compression via layer collapse operations, which efficiently explores the compression solution space through population-based search and a module-wise similarity fitness function that captures attention, feed-forward, and hidden state representations. We systematically evaluated compression at three different ratios (0.25, 0.50, and 0.75) and applied targeted post-training techniques to recover performance through fine-tuning and knowledge distillation over translation instructions. Additionally, we explored quantization techniques to achieve further model size reduction. Our experimental results demonstrate that the combination of evolutionary layer compression, targeted post-training, and quantization can achieve substantial model size reduction while maintaining competitive translation quality across all language pairs.

1 Introduction

The remarkable success of Large Language Models (LLMs) across diverse natural language processing tasks (Radford et al., 2019; Brown et al., 2020; Chang et al., 2024) has established them as powerful tools for language understanding and generation. Beyond their general capabilities, LLMs have also demonstrated remarkable effectiveness in machine translation tasks, often matching or exceeding the performance of dedicated neural machine translation systems (Xu et al., 2024; Zhu et al., 2024a; Kocmi et al., 2023, 2024). Simultaneously, recent work has focused on the development of specialized multilingual LLMs designed specifically for translation and cross-lingual tasks, such as Aya Expanse (Dang et al., 2024), EuroLLM (Martins et al., 2025), and Tower (Alves et al., 2024).

However, these advances come at the cost of substantial computational requirements. Modern LLMs, ranging from billions to trillions of parameters, demand considerable memory and processing power for both training and inference, with associated environmental impacts that raise serious sustainability concerns (Strubell et al., 2019). These computational requirements create barriers to widespread deployment and usage where reduced memory footprint and efficient inference are essential for practical adoption.

In this work, we describe Vicomtech's participation in the constrained track of the WMT 2025 Model Compression shared task (Gaido et al., 2025). This task focuses specifically on making LLMs suitable for deployment in machine translation within resource-constrained environments. The task evaluates compression techniques across multiple dimensions: model size reduction, translation quality preservation, and inference speed optimization. Participants were tasked to compress the Aya Expanse 8B model while maintaining competitive translation performance across three language pairs: Czech-German, English-Arabic, and Japanese-Chinese.

To address these challenges, we employed GeLaCo (Ponce et al., 2025), an evolutionary algorithm for LLM compression that builds upon the layer collapse operations of LaCo (Yang et al., 2024). GeLaCo efficiently explores the compression solution space through population-based search and a module-wise similarity fitness function that captures attention, feed-forward, and hidden state representations. We systematically applied this approach across multiple compression

Proceedings of the Tenth Conference on Machine Translation, pages 1011–1021
November 8-9, 2025

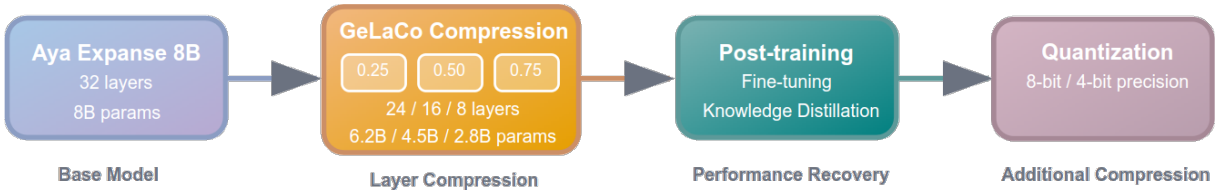


Figure 1: Overview of our model compression pipeline. Starting from Aya Expanse 8B (32 layers, 8B parameters), we apply GeLaCo compression at three ratios (0.25, 0.50, 0.75), reducing to 24, 16, or 8 layers respectively. Compressed models undergo post-training via Supervised Fine-Tuning or Generalized Knowledge Distillation for performance recovery, followed by optional 8-bit or 4-bit quantization for additional size reduction.

ratios (0.25, 0.50, and 0.75) for all three language pairs. We explored targeted post-training techniques, including continued pre-training and knowledge distillation, to recover translation performance after compression. Additionally, we used quantization methods to achieve further size reduction while maintaining translation quality.

Our experimental results demonstrate that the combination of evolutionary layer compression, targeted post-training, and quantization can achieve substantial model size reduction while preserving competitive translation capabilities across diverse language pairs. Figure 1 presents an overview of our compression pipeline.

2 Background

Model Compression. Traditional compression techniques for large language models include quantization, knowledge distillation, and pruning. Among pruning approaches, structured methods that remove entire layers or components have shown particular promise. Notable recent methods include SliceGPT (Ashkboos et al., 2024), which replaces sparse weight matrices with smaller dense matrices; LLM-Pruner (Ma et al., 2023), which uses gradient information to identify prunable components; and LaCo (Yang et al., 2024), which merges layers based on cosine similarity differences. However, these approaches typically require costly empirical evaluation of different compression schemes.

Evolutionary Compression. Evolutionary algorithms have recently emerged as a principled approach to explore the compression solution space. EvoPress (Sieberling et al., 2024) formulates compression as a general optimization problem using evolutionary algorithms for dynamic, non-uniform compression. DarwinLM (Tang et al., 2025) introduces training-aware structured pruning within an evolutionary framework. These methods demonstrate the potential of evolutionary approaches to discover better compression configurations compared to heuristic methods.

LLMs for Machine Translation. Large language models have become the dominant paradigm in machine translation, often matching or exceeding dedicated neural MT systems (Kocmi et al., 2023, 2024). This shift has motivated the development of specialized multilingual LLMs for translation tasks, such as Aya Expanse (Dang et al., 2024), which serves as the base model for our compression experiments. The success of LLMs in translation, combined with their substantial computational requirements, makes efficient compression particularly important for practical deployment in translation scenarios.

3 Methodology

3.1 GeLaCo

We employed GeLaCo, an evolutionary algorithm for LLM compression based on layer collapse operations. Layer collapse reduces model size by merging consecutive layers through differential weight merging, where the resulting parameters when merging m consecutive layers starting from layer l are computed as in Equation 1:

$$\theta_l^* = \theta_l + \sum_{k=1}^{m} (\theta_{l+k} - \theta_l), \tag{1}$$

where θ_l denotes the weight parameters of layer l, and $(\theta_{l+k} - \theta_l)$ denotes the parameter difference between each subsequent layer and the base layer l. This preserves contributions from collapsed layers while reducing the overall model size.

The main challenge in layer collapse lies in determining optimal merge operation combinations, as the search space grows exponentially with model size. Other approaches such as LaCo rely on empirical evaluation using heuristic methods,

which can be computationally expensive and may miss better compression solutions due to a limited exploration of the solution space.

GeLaCo addresses these limitations via population-based evolutionary search that efficiently explores compression configurations. The method uses a module-wise similarity fitness function that captures attention, feed-forward, and hidden state representations to guide the layer collapse operations through differential weight merging. The evolutionary process maintains a population of candidate solutions, where each individual represents a specific configuration of layer merge operations, evolving through fitness evaluation, selection, and crossover operations.

The fitness function evaluates compressed model quality by computing module-wise similarity between the original and compressed models using a small calibration dataset. For each calibration sentence, GeLaCo calculates cosine similarity across attention modules, feed-forward network components, and final hidden state representations, with the overall fitness score averaged across all three components and all calibration sentences. This approach enables an efficient evaluation of compression quality during the evolutionary search using only a small set of representative text samples.

3.2 Post-training

Previous work has demonstrated that instruction-following capabilities can be partially recovered through post-training of compressed models (Chen et al., 2025; Men et al., 2025; Ponce et al., 2025). We explored two distinct approaches for performance recovery. First, we applied Supervised Fine-Tuning (SFT) over translation instructions, where we adapted the compressed models to the translation task through continued training on parallel data. Alternatively, we explored Generalized Knowledge Distillation (GKD) (Agarwal et al., 2024), which addresses distribution mismatch by training the compressed student model on its own generated sequences while leveraging feedback from the original teacher model.

3.3 Quantization

To achieve further compression beyond layer collapse, we explored quantization techniques as a complementary approach. Quantization (Gray and Neuhoff, 1998) reduces the numerical precision of model parameters, offering additional size reductions while maintaining competitive performance (Zhu et al., 2024b). We systematically evaluated the combined effects of layer compression and quantization to understand their complementary potential for model size reduction.

3.4 In-context Learning

We investigated the effectiveness of in-context learning (ICL) to enhance the performance of compressed models across different prompting strategies. We explored three distinct setups: zero-shot translation, where we provided no examples; static few-shot learning, using a fixed set of translation examples; and retrieval augmented generation (RAG), using a dynamic similarity-based retrieval where we selected examples based on their relevance to the input sentence. This analysis allowed us to understand how compressed models respond to different contextual information and whether in-context learning can compensate for performance degradation from compression.

Dataset name	Total Size	Samples
CES-DEU		
Statmt-news_commentary-18.1	244,831	244,831
OPUS-neulab_tedtalks-v1	96,738	96,738
OPUS-ted2020-v1	153,227	153,227
OPUS-opensubtitles-v2024	36,408,370	168,401
OPUS-dgt-v4	3,048,670	168,401
OPUS-europarl-v8	568,589	168,402
Total		1,000,000
ENG-ARA		
OPUS-globalvoices-v2018q4	59,196	59,196
Statmt-news_commentary-18.1	193,671	193,671
Statmt-tedtalks-2_clean	341,887	149,426
OPUS-ted2020-v1	403,716	149,426
OPUS-qed-v2.0a	500,898	149,426
OPUS-opensubtitles-v2024	87,893,568	149,426
OPUS-multiun-v1	9,759,125	149,429
Total		1,000,000
JPN-ZHO		
Statmt-news_commentary-18.1	1,625	1,625
OPUS-ted2020-v1	15,982	15,982
Neulab-tedtalks_train-1	5,159	5,159
KECL-paracrawl-2wmt24	4,602,328	488,617
OPUS-opensubtitles-v2024	1,267,153	488,617
Total		1,000,000

Table 1: Dataset statistics for WMT 2025 Model Compression shared task training data. Total Size indicates the original dataset sizes, while Samples indicates the actual number of translation pairs used post-training.

Language Pair	Dataset	Samples
CES-DEU	newstests2019 (WMT 2024)	1,997
ENG-ARA	wmttest2024 (WMT 2024)	721
JPN-ZHO	WMT24++ (Deutsch et al., 2025)	998

Table 2: Test set statistics in terms of number of sentence pairs.

4 Experimental Setup

4.1 Models

Following the requirements of the constrained track of the shared task, we used the Aya Expanse 8B model as our foundation. This instruction-tuned model served both as our starting point for compression and as the primary baseline for performance comparison. We preserved the capabilities of the original model by not applying any additional training or modification to the base model prior to compression.

4.2 Corpora

Following the constrained track requirements, we sourced all training data from the WMT 2025 MT task data releases[1]. Our data selection strategy leveraged the available parallel corpora for each language pair, sampling from diverse sources of varying quality and domains. We arbitrarily selected one million translation instruction pairs per language combination as a compromise between coverage and reducing post-training computational time.

We provide a detailed breakdown of the original and sampled datasets in Table 1. Our training data consisted of one million translation instruction pairs for each of the three language pairs (Czech-German, English-Arabic, and Japanese-Chinese), yielding a total of 3 million translation instructions. We detail the specific instruction template used for translation post-training in Appendix B.

For evaluation, we selected test sets based on data released for WMT 2024[2]. Table 2 reports the number of translation pairs for each language pair and test set.

4.3 Compression

For the evolutionary search process, we used 16 randomly selected sentences from the monolingual portion of ParaCrawl for each target language, resulting in a total of 96 sentences as calibration data. We executed GeLaCo with the same configuration parameters as defined in the original work, running for 10,000 evolutionary steps with a single compression objective for each target ratio.

Using GeLaCo, we compressed the original 32-layer, 8-billion parameter model at three levels: 0.25 compression yielded 24 layers and approximately 6.2 billion parameters; 0.50 compression resulted in 16 layers and 4.5 billion parameters; and 0.75 compression produced 8 layers and 2.8 billion parameters.

For quantization, we employed the bitsandbytes library[3] to generate 8-bit and 4-bit with double quantization variants, providing additional compression beyond the structural layer reduction.

4.4 Post-training

We leveraged the 3 million translation instructions to perform both Supervised Fine-Tuning and Generalized Knowledge Distillation on the compressed models. For computational efficiency, we conducted all training using DeepSpeed with ZeRO Stage 3 optimization (Rajbhandari et al., 2020).

Due to the substantial computational requirements of GKD, we applied this technique exclusively to our smallest compressed model (0.75 compression ratio), while SFT was performed across all compression levels. The detailed hyperparameters for both SFT and GKD training, as well as the DeepSpeed ZeRO configuration, are provided in Appendix C.

4.5 Inference

We used vLLM (Kwon et al., 2023) for efficient inference across all experiments. For few-shot learning, we used 5 examples per evaluation. In the static few-shot setup, we randomly selected 5 translation instructions for each language pair from the training set. For dynamic few-shot learning, we performed BM25 retrieval over a subset of 10,000 training instances per language, selecting the 5 most similar translations to each source sentence. For retrieval, we used the Okapi BM25 implementation from Rank-BM25[4], configured with a minimum token length of 4 characters and whitespace tokenization.

[1] https://www2.statmt.org/wmt25/mtdata/
[2] https://data.statmt.org/wmt24/
[3] https://github.com/bitsandbytes-foundation/bitsandbytes
[4] https://github.com/dorianbrown/rank_bm25

Method	Size (GiB)	Inference	Time (s)	CES-DEU		ENG-ARA		JPN-ZHO	
				chrF	COMET	chrF	COMET	chrF	COMET
aya-expanse-8b	14.96	Zero-shot	88.67	54.1	0.8476	39.6	0.7699	23.5	0.8142
		Few-shot	60.32	53.7	0.8458	39.2	0.7709	22.5	0.8140
		RAG	619.14	53.5	0.8461	39.5	0.7721	23.0	0.8117
GeLaCo 0.25 - SFT	11.71	Zero-shot	78.36	51.2	0.8304	33.7	0.7327	17.8	0.7611
		Few-shot	83.24	51.2	0.8293	34.5	0.7300	18.0	0.7622
		RAG	616.42	51.3	0.8289	33.8	0.7200	17.7	0.7612
GeLaCo 0.50 - SFT	8.46	Zero-shot	73.61	48.3	0.7964	30.4	0.7105	13.0	0.7039
		Few-shot	66.74	48.1	0.7964	29.5	0.6976	12.3	0.69597
		RAG	641.88	48.2	0.7967	29.6	0.6945	12.0	0.6942
GeLaCo 0.75 - SFT	5.21	Zero-shot	62.98	40.7	0.6578	19.5	0.588	5.1	0.5499
		Few-shot	62.97	42.0	0.6612	18.9	0.5788	5.5	0.556
		RAG	588.99	40.2	0.6563	19.1	0.5797	5.2	0.5556
GeLaCo 0.75 - GKD	5.21	Zero-shot	49.79	50.3	0.7799	32.0	0.6964	16.7	0.7178
		Few-shot	53.14	14.5	0.3662	9.30	0.4271	1.3	0.4006
		RAG	440.86	13.4	0.3930	5.9	0.4377	1.2	0.3911

Table 3: Translation performance comparison across compression ratios. Results show chrF and COMET scores for zero-shot, few-shot, and RAG inference approaches across all three language pairs. SFT indicates fine-tuning and GKD indicates Generalized Knowledge Distillation post-training.

4.6 Evaluation

We evaluated translation quality using chrF (Popović, 2015) and COMET (Rei et al., 2020) with the Unbabel/wmt22-comet-da model[5]. For model size measurements, we report the VRAM usage during vLLM inference.

For inference speed evaluation, we report timing measurements computed using a batch size of 4,096, using bfloat16 precision for the non-quantized models. To ensure consistent timing comparisons, we excluded model loading times from our reported measurements due to the high variability and inconsistency that we observed during this phase and conducted 5 runs for each measurement to reduce variability and report the average results.

4.7 Hardware

For the GeLaCo compression process and inference experiments, we used a single NVIDIA L40 GPU with 48GB of vRAM. For post-training we used 4 NVIDIA H100 GPUs with 80GB of vRAM each for both SFT and GKD.

5 Results

We present our experimental results analyzing the impact of compression ratios, post-training approaches, and quantization on translation quality, model size, and inference speed.

5.1 Compression and Post-training

Table 3 presents the results for our baseline Aya Expanse 8B model and the compressed variants at compression ratios of 0.25, 0.50, and 0.75, along with their corresponding inference times for processing all three test sets. The results demonstrate the expected trade-off between model compression and translation quality across all evaluation settings.

As expected, model performance degrades progressively with increased compression levels, even after fine-tuning recovery. This pattern of declining performance with higher compression ratios is consistent across all language pairs and evaluation metrics. A notable result is that GKD demonstrates superior performance recovery compared to fine-tuning. For zero-shot translation, GKD at 0.75 compression shows improvements of +2.0, +1.6, and +3.7 chrF points for CES-DEU, ENG-ARA, and JPN-ZHO respectively compared to the 0.50 SFT model, while also showing an improvement of +1.4 COMET points in JPN-ZHO.

In-context Learning. Regarding in-context learning strategies, both static few-shot sampling and dynamic similarity-based retrieval yield results comparable to the zero-shot approach for the SFT models. However, the GKD-trained model presents a different behaviour, where ICL

[5]https://huggingface.co/Unbabel/wmt22-comet-da

Method	Quant.	Size (GiB)	CES-DEU		ENG-ARA		JPN-ZHO	
			chrF	COMET	chrF	COMET	chrF	COMET
aya-expanse-8b	-	14.96	54.1	0.8476	39.6	0.7699	23.5	0.8142
	Q8	8.46	54.2	0.8479	39.5	0.7661	23.5	0.8111
	Q4	5.31	53.9	0.8452	38.9	0.7661	22.8	0.8116
GeLaCo 0.25 - SFT	-	11.71	51.2	0.8304	33.7	0.7327	17.8	0.7611
	Q8	6.83	51.5	0.8318	33.7	0.7324	17.8	0.7606
	Q4	4.47	50.3	0.8242	32.1	0.7193	16.4	0.7477
GeLaCo 0.50 - SFT	-	8.46	48.3	0.7964	30.4	0.7105	13.0	0.7039
	Q8	5.21	48.2	0.7951	30.8	0.7125	12.7	0.6980
	Q4	3.63	46.5	0.7813	27.7	0.6878	12.3	0.6886
GeLaCo 0.75 - SFT	-	5.21	40.7	0.6578	19.5	0.588	5.1	0.5499
	Q8	3.58	40.3	0.6514	19.0	0.5782	4.9	0.5476
	Q4	2.79	35.3	0.6066	16.9	0.5491	4.5	0.5276
GeLaCo 0.75 - GKD	-	5.21	50.3	0.7799	32.0	0.6964	16.7	0.7178
	Q8	3.58	49.8	0.7789	32.5	0.6968	16.6	0.7204
	Q4	2.79	49.7	0.7756	31.1	0.6884	16.6	0.7174

Table 4: Impact of quantization on compressed model performance. Results compare 8-bit (Q8) and 4-bit (Q4) quantization using zero-shot translation across all language pairs.

methods fail dramatically. The GKD model's few-shot performance drops drastically across all language pairs, from 50.3 to 14.5 chrF for CES-DEU, 32.0 to 9.3 for ENG-ARA, and 16.7 to 1.3 for JPN-ZHO. Future work should address the drastic loss of in-context learning capabilities in GKD-trained models.

While we observe the expected reduction in processing time with model compression, from 88.67 seconds for the baseline to 49.79 seconds for the 0.75 GKD model in zero-shot setting, timing measurements showed unexpected variations where few-shot inference was occasionally faster than zero-shot despite the longer context, reflecting the inherent difficulties in timing measurements. Nevertheless, the computational overhead of the RAG approach consistently requires an order of magnitude more time, primarily due to the retrieval process overhead rather than the translation itself.

5.2 Quantization Results

Table 4 presents the results of applying quantization techniques to the GeLaCo compressed models, examining the effects of 8-bit and 4-bit quantization on model size and translation performance in a zero-shot setting.[6]

Across all GeLaCo variants, 8-bit quantization (Q8) reduces model sizes by approximately 42% while maintaining stable translation performance. For the 0.25 compressed model, Q8 reduces the size from 11.71 GiB to 6.83 GiB with minimal quality impact. The 0.50 compressed model follows a similar pattern, achieving a size reduction from 8.46 GiB to 5.21 GiB with marginal quality variations across language pairs. The 0.75 models also benefit from Q8 quantization, with both SFT and GKD variants reducing from 5.21 GiB to 3.58 GiB while preserving competitive performance.

4-bit quantization (Q4) enables more aggressive compression but introduces more noticeable quality degradation. For the 0.25 compressed model, Q4 reduces the size to 4.47 GiB while incurring chrF drops of 0.9, 1.6, and 1.4 points for CES-DEU, ENG-ARA, and JPN-ZHO respectively. This pattern intensifies with higher compression ratios, where the 0.75 SFT model with Q4 shows significant performance drops, particularly evident in the chrF scores falling to 35.3, 16.9, and 4.5.

A notable result emerges with the GKD variant, which demonstrates superior robustness to quantization. The 0.75 GKD model maintains competitive performance even with aggressive Q4 quantization, achieving chrF scores of 49.7, 31.1, and 16.6, substantially outperforming the corresponding SFT variant under the same quantization settings. The combination of layer compression and quantization enables the creation of extremely compact models, with the 0.75 GKD model reach-

[6]Complete quantization results including few-shot and RAG are provided in Appendix A

ing 2.79 GiB with Q4, representing an 81% reduction from the original baseline while retaining reasonable translation capabilities.

Given that quantization produces only marginal quality degradation while achieving substantial size reductions across all compression levels, we selected each of the Q8 and Q4 variants as our final submissions to the shared task. Specifically, we submitted the 0.25, 0.50, 0.75 SFT models and the 0.75 GKD model unquantized, Q8 and Q4 variants, with the 0.75 GKD Q4 model being designated as our primary submission due to its optimal balance of compression efficiency and translation quality preservation.

6 Conclusions

This work presented our approach to the WMT 2025 Model Compression shared task, focusing on compressing the Aya Expanse 8B model for machine translation across Czech-German, English-Arabic, and Japanese-Chinese language pairs within the constrained setting of the task. We employed GeLaCo, an evolutionary algorithm for layer collapse operations, combined with post-training techniques and quantization, to achieve substantial model size reduction while maintaining competitive translation performance.

Our experimental results demonstrated that compressed models can be successfully recovered through targeted post-training techniques. Generalized Knowledge Distillation consistently outperformed traditional fine-tuning for performance recovery across all three language pairs at the 0.75 compression ratio where it was applied. The combination of layer compression with 4-bit quantization achieved an 81% reduction in model size (from 14.96 GiB to 2.79 GiB) while preserving reasonable translation quality, making such models viable for resource-constrained scenarios.

Acknowledgments

This work was partially supported by the Department of Economic Development and Competitiveness of the Basque Government (SPRI Group) through funding for project IKUN (KK2024/00064).

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. SliceGPT: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*.

Tom Brown et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Yupeng Chang et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Xiaodong Chen, Yuxuan Hu, Jing Zhang, Yanling Wang, Cuiping Li, and Hong Chen. 2025. Streamlining redundant layers to compress large language models. In *The Thirteenth International Conference on Learning Representations*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker.

2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Marco Gaido, Thamme Gowda, Roman Grundkiewicz, and Matteo Negri. 2025. Findings of the WMT25 Model Compression Shared Task: Early Insights on Compressing LLMs for Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China.

R.M. Gray and D.L. Neuhoff. 1998. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-pruner: On the structural pruning of large language models. *Advances in Neural Information Processing Systems*, 36:21702–21720.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2025. Eurollm: Multilingual language models for europe. *Procedia Computer Science*, 255:53–62.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and weipeng chen. 2025. ShortGPT: Layers in large language models are more redundant than you expect.

David Ponce, Thierry Etchegoyhen, and Javier Del Ser. 2025. Gelaco: An evolutionary approach to layer compression. *arXiv preprint arXiv:2507.10059*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference*

for High Performance Computing, Networking, Storage and Analysis, pages 1–16. IEEE.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Oliver Sieberling, Denis Kuznedelev, Eldar Kurtic, and Dan Alistarh. 2024. EvoPress: Towards optimal dynamic model compression via evolutionary search. *arXiv preprint arXiv:2410.14649*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Shengkun Tang, Oliver Sieberling, Eldar Kurtic, Zhiqiang Shen, and Dan Alistarh. 2025. Darwinlm: Evolutionary structured pruning of large language models. *arXiv preprint arXiv:2502.07780*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Yifei Yang, Zouying Cao, and Hai Zhao. 2024. LaCo: Large language model pruning via layer collapse. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6401–6417, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024a. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024b. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.

A Full Results

Quant.	Method	Size (GiB)	Inference	CES-DEU		ENG-ARA		JPN-ZHO	
				chrF	COMET	chrF	COMET	chrF	COMET
-	aya-expanse-8b	14.96	Zero-shot	54.1	0.8476	39.6	0.7699	23.5	0.8142
			Few-shot	53.7	0.8458	39.2	0.7709	22.5	0.8140
			RAG	53.5	0.8461	39.5	0.7721	23.0	0.8117
	GeLaCo 0.25 - SFT	11.71	Zero-shot	51.2	0.8304	33.7	0.7327	17.8	0.7611
			Few-shot	51.2	0.8293	34.5	0.7300	18.0	0.7622
			RAG	51.3	0.8289	33.8	0.7200	17.7	0.7612
	GeLaCo 0.50 - SFT	8.46	Zero-shot	48.3	0.7964	30.4	0.7105	13.0	0.7039
			Few-shot	48.1	0.7964	29.5	0.6976	12.3	0.69597
			RAG	48.2	0.7967	29.6	0.6945	12.0	0.6942
	GeLaCo 0.75 - SFT	5.21	Zero-shot	40.7	0.6578	19.5	0.588	5.1	0.5499
			Few-shot	42.0	0.6612	18.9	0.5788	5.5	0.556
			RAG	40.2	0.6563	19.1	0.5797	5.2	0.5556
	GeLaCo 0.75 - GKD	5.21	Zero-shot	50.3	0.7799	32.0	0.6964	16.7	0.7178
			Few-shot	14.5	0.3662	9.30	0.4271	1.3	0.4006
			RAG	13.4	0.3930	5.9	0.4377	1.2	0.3911
Q8	aya-expanse-8b	14.96	Zero-shot	54.2	0.8479	39.5	0.7661	23.5	0.8111
			Few-shot	53.9	0.8472	39.3	0.7682	22.5	0.8153
			RAG	53.7	0.8466	39.4	0.7713	22.9	0.8113
	GeLaCo 0.25 - SFT	11.71	Zero-shot	51.5	0.8318	33.7	0.7324	17.8	0.7606
			Few-shot	51.5	0.8316	34.6	0.7304	18.1	0.7641
			RAG	51.4	0.8299	33.2	0.7205	18.0	0.7616
	GeLaCo 0.50 - SFT	8.46	Zero-shot	48.2	0.7951	30.8	0.7125	12.7	0.6980
			Few-shot	48.6	0.7953	29.7	0.6983	12.4	0.6925
			RAG	48.9	0.7976	29.8	0.6928	12.2	0.6968
	GeLaCo 0.75 - SFT	5.21	Zero-shot	40.3	0.6514	19.0	0.5782	4.9	0.5476
			Few-shot	39.8	0.6516	17.6	0.5678	5.4	0.5546
			RAG	40.3	0.6537	17.3	0.5692	5.0	0.5472
	GeLaCo 0.75 - GKD	5.21	Zero-shot	49.8	0.7789	32.5	0.6968	16.6	0.7204
			Few-shot	14.2	0.3732	10.1	0.4374	1.4	0.4049
			RAG	13.0	0.3923	6.1	0.4378	1.3	0.3948
Q4	aya-expanse-8b	14.96	Zero-shot	53.9	0.8452	38.9	0.7661	22.8	0.8116
			Few-shot	53.2	0.8431	39.1	0.7681	21.7	0.8140
			RAG	53.2	0.8425	39.1	0.7685	22.4	0.8076
	GeLaCo 0.25 - SFT	11.71	Zero-shot	50.3	0.8242	32.1	0.7193	16.4	0.7477
			Few-shot	50.3	0.8239	32.7	0.7155	16.7	0.7477
			RAG	50.4	0.8238	31.6	0.7076	16.6	0.7496
	GeLaCo 0.50 - SFT	8.46	Zero-shot	46.5	0.7813	27.7	0.6878	12.3	0.6886
			Few-shot	47.3	0.7807	28.8	0.6801	12.4	0.6845
			RAG	47.4	0.7831	27.8	0.6733	12.5	0.6861
	GeLaCo 0.75 - SFT	5.21	Zero-shot	35.3	0.6066	16.9	0.5491	4.5	0.5276
			Few-shot	37.0	0.6164	16.9	0.5571	4.5	0.5311
			RAG	37.3	0.6196	16.6	0.552	4.5	0.5353
	GeLaCo 0.75 - GKD	5.21	Zero-shot	49.7	0.7756	31.1	0.6884	16.6	0.7174
			Few-shot	12.9	0.3767	9.9	0.3931	1.4	0.4009
			RAG	12.3	0.3874	5.9	0.4277	1.4	0.3949

Table 5: Complete experimental results across all models, quantization settings, and inference approaches.

B Translation Instruction Template

The following template illustrates the format used for translation instructions. At inference time, only the user message is provided to the model. The instruction prompt and language names are always specified in English. In the template, SOURCE_LANGUAGE and TARGET_LANGUAGE represent the English names of the source and target languages (e.g., "Czech", "German", "English", "Arabic", "Japanese", or "Chinese"), INPUT_SENTENCE contains the text to be translated in the source language, and TARGET_SENTENCE contains the corresponding translation in the target language.

Instruction Template

```
"messages": [
    {
        "role": "user",
        "content": "Translate from SOURCE_LANGUAGE to TARGET_LANGUAGE:\nINPUT_SENTENCE"
    },
    {
        "role": "assistant",
        "content": "TARGET_SENTENCE"
    }
]
```

C Training Hyperparameters

For both supervised fine-tuning and generalized knowledge distillation, we employed the SFTTrainer and GKDTrainer implementations from the TRL[7] library. All training was conducted using DeepSpeed with ZeRO Stage 3 optimization for efficient memory management across multiple GPUs. The specific hyperparameters used for each training approach are detailed below.

SFT Training Hyperparameters

```
--learning_rate 2.0e-5
--num_train_epochs 3
--packing
--per_device_train_batch_size 8
--gradient_accumulation_steps 4
--gradient_checkpointing
--bf16 True
```

GKD Training Hyperparameters

```
--learning_rate 2.0e-5
--per_device_train_batch_size 4
--gradient_accumulation_steps 8
--bf16 True
--logging_steps 25
```

DeepSpeed ZeRO Configuration

```
compute_environment: LOCAL_MACHINE
debug: false
deepspeed_config:
  deepspeed_multinode_launcher: standard
  offload_optimizer_device: none
  offload_param_device: none
  zero3_init_flag: true
  zero3_save_16bit_model: true
  zero_stage: 3
distributed_type: DEEPSPEED
downcast_bf16: 'no'
machine_rank: 0
main_training_function: main
mixed_precision: bf16
num_machines: 1
num_processes: 8
rdzv_backend: static
same_network: true
tpu_env: []
tpu_use_cluster: false
tpu_use_sudo: false
use_cpu: false
```

[7] https://huggingface.co/docs/trl/index

Iterative Layer Pruning for Efficient Translation Inference

Yasmin Moslem*
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
yasmin.moslem@adaptcentre.ie

Muhammad Hazim Al Farouq*
Kreasof AI
Research Labs
Jakarta, Indonesia
muhammad.hazim@kreasof.my.id

John D. Kelleher
ADAPT Centre
Trinity College Dublin
Dublin, Ireland
john.kelleher@adaptcentre.ie

Abstract

Large language models (LLMs) have transformed many areas of natural language processing, including machine translation. However, efficient deployment of LLMs remains challenging due to their intensive computational requirements. In this paper, we address this challenge and present our submissions to the *Model Compression* track at the Conference on Machine Translation (WMT 2025). In our experiments, we investigate iterative layer pruning guided by layer importance analysis. We evaluate this method using the Aya-Expanse-8B model for translation from Czech to German, and from English to Egyptian Arabic. Our approach achieves substantial reductions in model size and inference time, while maintaining the translation quality of the baseline models.

1 Introduction

Large language models (LLMs) have demonstrated powerful capabilities in diverse natural language processing tasks, including translation. However, LLMs are often computationally intensive, making them impractical to deploy in real-world settings with limited resources. To enhance the efficiency of these models, researchers have explored various model compression techniques, aiming to reduce their computational requirements while preserving quality (Gandhi et al., 2023; Sajjad et al., 2023; Treviso et al., 2023; Sreenivas et al., 2024; Gu et al., 2025; Moslem, 2025).

Aya Expanse is an open-weight large language model with multilingual capabilities. The WMT 2025 Model Compression track (Gaido et al., 2025) required all submissions to be derived from the Aya-Expanse-8B model. This work focuses on translation from Czech to German and from English to Egyptian Arabic.

Our experiments build on established work on iterative layer pruning guided by layer importance evaluation (Peer et al., 2022; Moslem, 2025). We apply iterative layer pruning to the baseline model Aya-Expanse-8B[1] which originally consists of 32 layers and 8.03B parameters. This approach incrementally identifies and removes layers with minimal contribution to translation quality, one layer at a time. To this end, we conduct layer importance evaluation by measuring translation performance without each layer. After identifying and removing the least critical layer, we repeat the layer importance evaluation on the remaining layers until reaching our pruning target. The pruned model resulting from this process is then fine-tuned on the News Commentary dataset. We have made three submissions; the primary submission is a 24-layer model with 6.28B parameters, and the two contrastive submissions are 20-layer and 16-layer models, with 5.41B and 4.54B parameters, respectively.

2 Data

After layer pruning of the *Aya-Expanse-8B* model (cf. Section 3), we need to fine-tune the pruned model on medium-sized training data to restore the translation quality of the baseline model. To this end, we use the News Commentary dataset[2] which consists of news articles and their corresponding translations in several languages, including Arabic, English, German, and Czech.

We start by rule-based filtering of the Czech-to-German (CES-DEU) News Commentary dataset by removing duplicates, segments longer than 200 words, and those whose source/target length ratio is larger than 1.5 times. We also apply language detection with fastText[3] (Joulin et al., 2017) with a 0.9 threshold. Finally, we conduct semantic filtering using the *mUSE* model (Yang et al.,

*These authors contributed equally to this work.

[1] https://hf.co/CohereLabs/aya-expanse-8b

[2] https://data.statmt.org/news-commentary/v18/training/

[3] In particular, we used the fastText *"lid.176.bin"* model.

Proceedings of the Tenth Conference on Machine Translation, pages 1022–1027
November 8-9, 2025

2020) and *Sentence-Transformers* (Reimers and Gurevych, 2019) with a 0.7 threshold of semantic similarity between the source and target. The CES-DEU News Commentary dataset includes 250.4K segments before filtering, and 201.3K segments after filtering.[4] Eventually, we split the dataset into train and test splits, where the test set includes 500 segments used for both testing and layer importance evaluation. Then, we sample 100K of the training data, using 0 as the random seed in both cases.

As the English-to-Arabic News Commentary dataset uses Standard Arabic,[5] we first apply the same rule-based and semantic filtering steps as those we employ while processing the Czech-to-German dataset, which result in 84.3K segments. Afterwards, we convert Standard Arabic text segments into Egyptian Arabic (ARZ) with GPT-4.1-Mini, using the prompt in Appendix A, providing a fixed verified example that includes the English source as well as both the Standard Arabic and Egyptian Arabic translations. For parameters, we use temperature 0.3 and top-p 1. After completing the generation of the synthetic Egyptian Arabic translations, we apply rule-based filtering, comparing the generated Egyptian Arabic text segments to the original English source. Finally, we calculate the semantic similarity between the English source and the Egyptian Arabic target and select the 500 segments with the highest scores (0.91-0.98) for the "test" split, while the "train" split comprises the remaining 83.2K segments.[6] Using this synthetic dataset to fine-tune our models yields clear quality gains compared to the baseline models, when evaluated on both the in-domain holdout test dataset (cf. Table 1) and the WMT24++ benchmark[7] (Deutsch et al., 2025) which includes 998 segments (cf. Table 3).

3 Iterative Layer Pruning

As previous research demonstrates, iterative layer pruning achieves better quality than middle layer pruning (Moslem, 2025). In this experimental setup, we apply iterative layer pruning to the *Aya-Expanse-8B* baseline model. This approach incrementally identifies and removes layers with minimal contribution to translation quality, one layer at a time. The pruned models resulting from this process are then fine-tuned on the training dataset. Furthermore, knowledge distillation data from the teacher model can be added. Fine-tuning the pruned model restores most of the baseline model's translation quality. The following points elaborate on the process.

Layer importance evaluation: We conduct layer importance evaluation by measuring translation performance without each layer. In this greedy layer pruning approach (Peer et al., 2022; Rostami and Dousti, 2024; Moslem, 2025), to prune $n + 1$ layers, only a single optimal layer to prune must be added to the already known solution for pruning n layers. After identifying and removing the least critical layer, we repeat the layer importance evaluation on the remaining layers until reaching our n pruning target. We observe that while removing certain layers of the model (e.g. the first or last layer) substantially degrades translation performance, others result in minimal performance drops. Following Moslem (2025), we use the chrF++ metric for layer importance evaluation for both better efficiency and quality.

Layer pruning: We iteratively prune one decoder layer at a time, selecting the layer whose removal has the least negative impact on translation quality, measured by chrF++ scores. At each iteration, we evaluate the translation performance of the pruned model on the test split of the News Commentary dataset, after removing each candidate layer. The layer whose removal yields the best performance is eventually pruned. This process continues until a predefined number of layers (8, 12, and 16 layers) have been removed. By iteratively removing the least important layers, this performance-guided method produces a more compact model that can be fine-tuned further to recover the translation quality of the original model. We observe that the performance of the CES-DEU model is more impacted by pruning than the ENG-ARZ model, which might be attributed to the pre-training process (cf. Table 1). In other words, the evaluation of the baseline for CES-DEU translation achieves better results than that for ENG-ARZ translation; hence, it seems that fine-tuning the pruned ENG-ARZ models has helped with improving the translation quality of this language pair.

Fine-tuning: The pruning step is followed by fine-tuning the pruned model for 1 epoch using the News Commentary dataset (cf. Section 2). The

[4] https://hf.co/datasets/ymoslem/news-commentary-cs-de
[5] https://hf.co/datasets/ymoslem/news-commentary-en-ar
[6] https://hf.co/datasets/ymoslem/news-commentary-eng-arz
[7] https://hf.co/datasets/google/wmt24pp

Language	Model	Layers	chrF++ ↑	COMET ↑	Params (B) ↓	Speed (mm:ss) ↓
CES-DEU	Baseline	32	**52.79**	**87.18**	8.03	00:47
	Pruned + FT	24	51.35	85.70	6.28	00:34
		20	49.45	83.95	5.41	**00:27**
		16	45.79	79.39	**4.54**	**00:27**
ENG-ARZ	Baseline	32	42.03	81.45	8.03	01:22
	Pruned + FT	24	**58.38**	**85.74**	6.28	00:54
		20	55.69	84.50	5.41	00:51
		16	51.17	82.10	**4.54**	**00:42**

Table 1: Evaluation of layer pruning experiments. For translation from Czech to German (CES-DEU), pruning 8 layers and then fine-tuning the resulting model retains 98% of the translation quality (as measured by COMET). Interestingly, for translation from English to Egyptian Arabic (ENG-ARZ), the model resulting from pruning up to 16 layers and then fine-tuning outperforms the Aya-Expanse-8B baseline for this language pair.

training uses a learning rate of 2e-5, a batch size of 8, and early stopping with a patience value of 5 evaluation runs, and it is conducted on one A100 80GB GPU. This fine-tuning step recovers most of the translation quality of the baseline model.

Model	Layers	KD	chrF++ ↑	COMET ↑
Baseline 32B	40	-	54.57	87.76
Baseline 8B	32	-	52.79	87.18
Pruned + FT	24	⊗	51.35	85.70
		✓	52.68	86.50
	20	⊗	49.45	83.95
		✓	51.25	85.19
	16	⊗	45.79	79.39
		✓	48.60	81.39

Table 2: Evaluation of knowledge distillation (KD). Fine-tuning pruned models on a combination of authentic and synthetic data (generated by Aya-Expanse-32B) improved the CES-DEU translation quality, with the 24-layer pruned model nearly matching the performance of the Aya-Expanse-8B baseline.

Knowledge distillation: To improve the quality of the CES-DEU models, we employed sequence-level knowledge distillation, where the student model is fine-tuned on a combination of authentic data and synthetic data generated by the teacher model for the same training dataset. In this case, the teacher model is the Aya-Expanse-32B while the students are the pruned models. After generating the data, we filter it by removing duplicates (exact matches in the target side of the authentic data), and translations with less than 70% COMET scores, resulting in extra 98.6K segments of training data (cf. Section 2). As Table 2 demonstrates, fine-tuning the pruned models with a combination of both the authentic and knowledge distillation data has improved their translation quality, and helped close the performance gap between the 24-layer pruned model and the Aya-Expanse-8B baseline. Similarly, the 20-layer and 16-layer models show 2-3 points of improvement in terms of chrF++ and COMET metrics.

4 Inference and Evaluation

For inference, we use greedy generation by disabling the sampling options, and setting the temperature argument to 0. We apply a simple translation prompt: "Translate the following text from {source_language} to {target_language}:"

To evaluate our systems, we calculated BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), as implemented in the sacreBLEU library[8] (Post, 2018). For semantic evaluation, we use COMET (Rei et al., 2020).[9] Table 1 reports the results of the main experiments using the *Transformers* framework[10] (Wolf et al., 2020) for inference.

5 Results

The process of iterative layer pruning has achieved model compression from 8.03B parameters to 6.28B, 5.41B, and 4.54B parameters, after removing 8, 12, and 16 layers, respectively. Moreover, the quality degradation caused by pruning has been mitigated through fine-tuning on medium-sized data

[8] https://github.com/mjpost/sacrebleu

[9] In particular, we used the *"wmt22-comet-da"* model.

[10] https://github.com/huggingface/transformers

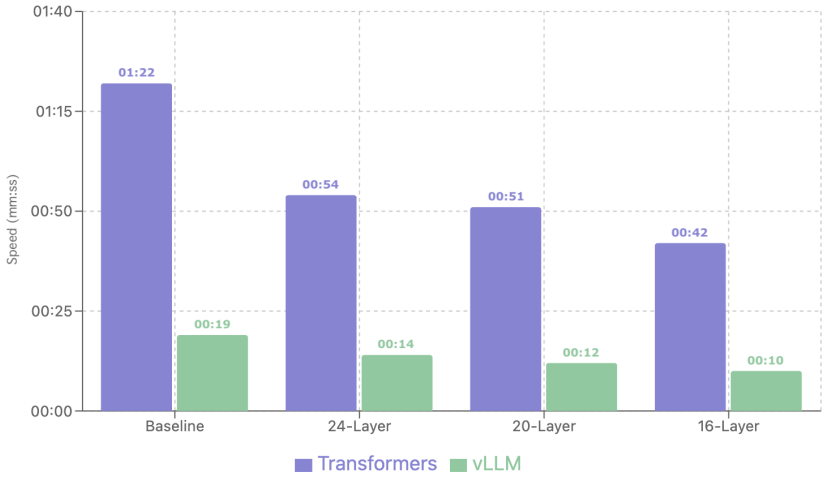


Figure 1: Inference speed comparison between *Transformers* and *vLLM*, using the Aya-Expanse-8B model for ENG-ARZ translation. *vLLM* consistently outperforms *Transformers* across all model sizes. Speedup ranges from 4.2x (16-layer) to 4.3x (baseline model). Both frameworks show improved performance with layer pruning. The 16-layer model achieves the fastest inference times overall.

(80K-100K) and knowledge distillation. As demonstrated by Table 1, by the end of the process, the pruned model could recover most of the translation quality of the baseline model. For translation from Czech to German (CES-DEU), pruning 8 layers and then fine-tuning the resulting model retains 98% of the translation quality (as measured by COMET) before knowledge distillation and 99% after knowledge distillation. Interestingly, for translation from English to Egyptian Arabic (ENG-ARZ), the model resulting from pruning up to 16 layers and then fine-tuning outperforms the baseline model. This can be attributed to the initial quality of the baseline model for this language pair.

Moreover, we experimented with immediate recovery through fine-tuning the model after each pruning phase (i.e. pruning the fine-tuned 24-layer model into 20 layers instead of pruning the baseline model directly), and noticed that the final quality was similar to pruning the baseline directly and then only fine-tuning the pruned model. This matches the results demonstrated by Moslem (2025) who experimented with immediate fine-tuning after pruning each layer, and observed that this could lead to overfitting. In other words, it is sufficient to fine-tune the final pruned model.

In terms of inference performance, we observe that using *vLLM* [11] (Kwon et al., 2023) as an inference engine instead of *Transformers* increases the inference speed by more than four times when conducting the evaluation on one A40 48GB GPU (cf. Figure 1). Moreover, while 4-bit quantization using *bitsandbytes* (Dettmers et al., 2023) reduces the memory footprint, pruning results in higher inference speed and throughput (cf. Table 4).

6 Conclusions and Future Work

In this work, we demonstrated that iterative layer pruning is an effective approach for compressing LLMs while retaining translation quality. The method relies on layer importance evaluation, followed by fine-tuning on a medium-sized dataset. This iterative layer pruning process reduces the model size and accelerates inference. To ensure reproducibility, we have made our code publicly available.[12]

Future research directions include investigating adaptive compression approaches that dynamically select appropriate model configurations based on real-time deployment constraints such as memory limits and latency requirements. Moreover, we plan to assess our compression methods on a broader range of datasets, including both sentence-level and document-level data. Since Aya-Expanse is designed to follow textual instructions, exploring retrieval-augmented generation combined with few-shot prompting presents a promising opportunity for enhancing translation performance in compressed models.

[11] https://github.com/vllm-project/vllm

[12] https://github.com/ymoslem/Model-Compression

Acknowledgements

We sincerely thank the ADAPT Centre (Ireland) and Kreasof AI (Indonesia) for providing the resources and support that made this work possible.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv [cs.LG]*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, and 14 others. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects. *arXiv preprint arXiv:2502.12404*.

Marco Gaido, Thamme Gowda, Roman Grundkiewicz, and Matteo Negri. 2025. Findings of the WMT25 Model Compression Shared Task: Early Insights on Compressing LLMs for Machine Translation. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv [cs.CL]*.

Yuxian Gu, Qinghao Hu, Shang Yang, and 4 others. 2025. Jet-Nemotron: Efficient language model with Post Neural Architecture Search. *arXiv [cs.CL]*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, and 6 others. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, New York, NY, USA. ACM.

Yasmin Moslem. 2025. Efficient speech translation through model compression and knowledge distillation. In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 379–388, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

David Peer, Sebastian Stabinger, Stefan Engl, and Antonio Rodríguez-Sánchez. 2022. Greedy-layer pruning: Speeding up transformer models for natural language processing. *Pattern Recognit. Lett.*, 157:76–82.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Pedram Rostami and Mohammad Javad Dousti. 2024. CULL-MT: Compression using language and layer pruning for machine translation. *arXiv [cs.CL]*.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Comput. Speech Lang.*, 77(101429):101429.

Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, and 19 others. 2024. LLM pruning and distillation in practice: The Minitron approach. *arXiv [cs.CL]*.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, and 19 others. 2023. Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Thomas Wolf, Lysandre Debut, Victor Sanh, and 19 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, and 9 others. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

A Prompt for Synthetic Data Generation for Egyptian Arabic

```
I would like to convert a Standard Arabic text into Egyptian Arabic.
Please generate the Egyptian Arabic version using a neutral, informative
tone with slightly conversational phrasing, similar to the example below.
The output should feel natural, like it's written for a general Egyptian
audience but still accurate and clear. Do not add any commentary; just
return the Egyptian Arabic version.

English:
<english_example>

Standard Arabic:
<standard_arabic_example>

Egyptian Arabic:
<egyptian_arabic_example>

English:
{new_source_text}

Standard Arabic:
{new_target_text}

Egyptian Arabic:
```

B Evaluation of Egyptian Arabic Translation on WMT24++

Model	Layers	chrF++ ↑	COMET ↑
Baseline 32B	40	33.89	75.55
Baseline 8B	32	30.62	74.50
Pruned + FT	24	**37.01**	**76.86**
	20	34.24	74.95
	16	29.32	68.70

Table 3: Evaluation of the ENG-ARZ models fine-tuned with target-side synthetic data. The evaluation uses the WMT24++ benchmark and shows quality improvement compared to the baseline models.

C Quantization Speed and Throughput

Model	Layers	4-bit	Memory ↓	Speed ↓	Throughput ↑
Baseline 8B	32	no	14.96	00:19	2275
		yes	5.61	00:42	1053
Pruned + FT	24	no	11.71	00:14	3008
		yes	4.70	00:22	2004
	20	no	10.08	00:12	3484
		yes	4.24	00:18	2367
	16	no	8.46	00:10	4192
		yes	3.78	00:15	2908

Table 4: Performance comparison of Aya-Expanse-8B baseline and the pruned models with and without 4-bit quantization, in terms of memory (GiB), speed (mm:ss), and output throughput (tokens/sec). The evaluation uses the holdout ENG-ARZ News Commentary test dataset, on one A40 48GB GPU. While 4-bit quantization reduces the memory footprint, layer pruning achieves both higher inference speed and throughput.

Expanding the WMT24++ Benchmark with Rumantsch Grischun, Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader

Jannis Vamvas[1] **Ignacio Pérez Prat**[2] **Not Battesta Soliva**[1]
Sandra Baltermia-Guetg[2] **Andrina Beeli**[2] **Simona Beeli**[2] **Madlaina Capeder**[2]
Laura Decurtins[2] **Gian Peder Gregori**[2] **Flavia Hobi**[2] **Gabriela Holderegger**[2]
Arina Lazzarini[2] **Viviana Lazzarini**[2] **Walter Rosselli**[2] **Bettina Vital**[2]
Anna Rutkiewicz[1] **Rico Sennrich**[1]

[1]University of Zurich [2]Lia Rumantscha

Correspondence: vamvas@cl.uzh.ch, ignacio.perez.prat@rumantsch.ch

Abstract

The Romansh language, spoken in Switzerland, has limited resources for machine translation evaluation. In this paper, we present a benchmark for six varieties of Romansh: Rumantsch Grischun, a supra-regional variety, and five regional varieties: Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. Our reference translations were created by human translators based on the WMT24++ benchmark, which ensures parallelism with more than 55 other languages. An automatic evaluation of existing MT systems and LLMs shows that translation out of Romansh into German is handled relatively well for all the varieties, but translation into Romansh is still challenging.

1 Introduction

The automatic evaluation of machine translation (MT) has been widened in recent years to cover more languages and language varieties. While massively multilingual benchmarks such as FLORES (Goyal et al., 2022; NLLB Team et al., 2024) or NTREX (Federmann et al., 2022) include reference translations in hundreds of languages, no dedicated reference translations for the Romansh language have been available so far. In this paper, we close this gap by extending the recent WMT24++ benchmark (Kocmi et al., 2024; Deutsch et al., 2025) with reference translations for six varieties of Romansh, using German as the source language.

There are several reasons why Romansh, which is a language from the Romance family spoken in Switzerland (**ISO 639-1: `rm`; ISO 639-2/3: `roh`**), has had limited resources for MT evaluation. First, Romansh is considered a minority language, with 40,000–60,000 speakers (Gross, 2004; Grünert, 2024). Secondly, multiple varieties of Romansh need to be considered for a comprehensive evaluation. *Rumantsch Grischun* is a supra-regional variety of the language, often used in official contexts. However, the five regional varieties of Romansh (Figure 1), usually referred to as *idioms*, are more widely spoken in everyday life, with limited mutual intelligibility (Gross, 2004).

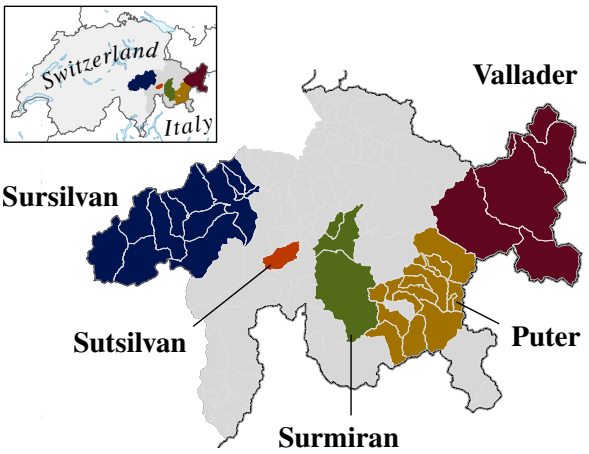


Figure 1: Distribution of Romansh *idioms* (regional varieties) within south-eastern Switzerland. The map shows municipalities where an idiom is officially used in public administration. We extend the WMT24++ benchmark with sets of reference translations for these five idioms, as well as Rumantsch Grischun, a supra-regional variety of Romansh.

Prior work on MT for Romansh (Müller et al., 2020; Niklaus et al., 2025) has leveraged multilingual government press releases (Scherrer and Cartoni, 2012), blog posts or federal laws, all of which cover only the Rumantsch Grischun variety. Our benchmark based on WMT24++ enables a more systematic evaluation setup that includes the five idioms and a broader range of domains, such as social media and transcripts of YouTube videos.

We release our benchmark under the Apache 2.0 license.[1] In addition, we use the benchmark to perform a systematic evaluation of MT systems and LLMs on German–Romansh and Romansh–German translation. Results based on automatic evaluation metrics indicate that translation into German achieves reasonable quality for all Romansh

[1] https://hf.co/datasets/ZurichNLP/wmt24pp-rm

Proceedings of the Tenth Conference on Machine Translation, pages 1028–1047
November 8-9, 2025

	Data Sample
English (Kocmi et al., 2024)	*it seems like even iMessage over WiFi isn't working, which doesn't quite make sense to me*
German (Deutsch et al., 2025)	*Anscheinend funktioniert nicht mal iMessage über WiFi, was mir nicht ganz einleuchtet*
Rumantsch Grischun Code: `roh_Latn_ruma1247`	*Para che gnanc iMessage funcziunia via WiFi, tge ch'è per mai betg dal tut evident.*
Sursilvan Code: `roh_Latn_surs1244`	*Sco ei para funcziunescha gnanc iMessage sur WiFi, quei ch'jeu sai buca propi capir.*
Surmiran Code: `roh_Latn_surm1243`	*Scu para funcziunescha mianc iMessage sur WiFi, chegl tg'ia sa betg propi tgapeir*
Sutsilvan Code: `roh_Latn_suts1235`	*Para funcziunescha gnànc iMessage sur igl WiFi, tge ca fa betga propi sen tanor me*
Puter Code: `roh_Latn_uppe1396`	*Pera cha nu funcziuna niauncha iMessage sur WiFi, che ch'eau nun incleg dal tuot*
Vallader Code: `roh_Latn_lowe1386`	*Apparaintamaing nu funcziuna gnanca üna jada iMessage sur WiFi, quai chi nu'm voul propcha ir per testa*

Table 1: Samples of the Romansh varieties that we contribute to the benchmark, plus the English and German segments from prior work. The samples are from the *Social* domain. The language code assigned by the Open Language Data Initiative (OLDI) has three components: the ISO 639-3 language code (`roh`), the ISO 15924 script code (`Latn`), and the Glottocode assigned to the variety by Glottolog (Hammarström et al., 2025).

varieties, while translation into Romansh remains challenging, particularly for the less-resourced idioms. Code for reproducing our experiments is available.[2]

2 Language Overview

2.1 Romansh

Romansh is part of the Romance branch of the Indo-European language family. It is a minority language in the Swiss canton of Graubünden and is treated as one of the country's four national languages (Grünert, 2018). Its status is considered endangered (Moseley and Nicolas, 2010). Romansh covers an extremely diverse dialect continuum spanning the canton, where roughly 15% of inhabitants speak it as their main language (Gross, 2004). Unlike other dialect continuums, Romansh is not "roofed" by a single standard language (Goebl, 2003). Instead, there are five different written traditions dividing the Romansh-speaking area into regions with their own written standards—known as *idioms*—that differ heavily from each other in all areas of language structure (Liver, 2010; Haiman and Benincà, 1992; Schmid, 1976). In the 1980s, linguist Heinrich Schmid developed Rumantsch Grischun, a supra-regional, constructed standard, as a *Dachsprache* for Romansh (Muljačić, 2012).

2.2 Sursilvan

Sursilvan is used in the west of Graubünden, in an area mainly covered by the Surselva valley. In many municipalities towards the east, it is still the predominant first language, while German is increasingly dominant moving west (Gross, 2004). Sursilvan is the idiom with the largest population. The written form mainly represents the dialects spoken between Disentis and Ilanz, though the entire area it covers is a continuum exhibiting mutual intelligibility. There were at least 18,000 Sursilvan speakers in the year 2000 (Gross, 2004).

2.3 Sutsilvan

Sutsilvan is spoken in the valley of the Hinterrhein river, though its territory is no longer contiguous. It is the variety with the lowest number of speakers and the highest level of endangerment (Liver, 2014). Large parts of Sutsilvan's traditional speaker territory became German speaking several centuries ago, and until recently, there was no established written form for Sutsilvan. A concentrated

[2] https://github.com/ZurichNLP/romansh_mt_eval

effort to change this was initiated by Giuseppe Gangale in the 1940s, establishing modern Sutsilvan orthography. His approach, however, sparked debate, and Sutsilvan remains an idiom with hardly any majority Romansh territory (Coray, 2008). There were at least 1,000 Sutsilvan speakers in the year 2000 (Gross, 2004).

2.4 Surmiran

Surmiran is spoken in central Graubünden, namely in the regions of Alvra/Sotses and Surses, the latter being an area where Romansh is still largely present in everyday life (Liver, 2014). Together with Sutsilvan, Surmiran constitutes a bridge between the starkly different dialects of the Surselva region in western Graubünden and the Engadine valley in the east. For this reason, it has previously been suggested as a lingua franca for supra-regional communication (Coray, 2008). Surmiran itself exhibits some peculiarities, however, shared by neither of the two other major Romansh-speaking areas. There were at least 3,000 Surmiran speakers in the year 2000 (Gross, 2004).

2.5 Puter

Puter and Vallader are used as written standards in the Engadine valley, with Puter being used south of Zuoz. The Engadine valley can itself be seen as a continuum of varieties more diverse than the Surselva (Schmid, 1976). Puter reflects characteristics of the dialects in the upper Engadine valley, with more Italian influence than Vallader. Written Puter dates back to 1552 (Obrist, 2022), and is thus the variety with the longest-standing written tradition. Puter is under substantial pressure from German due to growing tourism since the last century (Liver, 2014). Municipalities with a Romansh majority have become scarce. There were at least 5,500 Puter speakers in the year 2000 (Gross, 2004).

2.6 Vallader

Vallader is used in the Lower Engadine valley, north of Zernez, as well as in the Val Müstair. Vallader, unlike Puter, remains a majority language in most of its territory (Liver, 2014). Together with Sursilvan, the Vallader territory represents a stronghold of Romansh. There were at least 6,500 Vallader speakers in the year 2000 (Gross, 2004).

2.7 Rumantsch Grischun

The Rumantsch Grischun variety has a special role in that it is not an idiom, but a written standard devised as a constructed language. It does not reflect any Romansh speaker's natural speech, but was constructed to be a globally intelligible and neutral written form that could be used to represent Romansh as a language. It was developed by comparing structural and lexical characteristics of the different idioms and determining the most mutually intelligible forms (Schmid, 1982).

Rumantsch Grischun is used for official publications from the canton or the federal government, as well as other institutions addressing the entire Romansh population. More extended promotion of Rumantsch Grischun (including replacing the idioms as the language of literacy at schools) met heavy resistance and caused long-lasting debate (Coray, 2008). Most speakers of Romansh only actively learn their own idiom. Though they may occasionally come into contact with Rumantsch Grischun texts, their knowledge of it is only passive at most.

3 Data Collection

3.1 Choice of Benchmark

We chose to extend the WMT24++ benchmark (Kocmi et al., 2024; Deutsch et al., 2025) based on the following considerations:

- WMT24++ currently covers 55 languages, including the other Swiss national languages (German, French and Italian).
- It is a recent benchmark that is unlikely to suffer from data contamination in LLMs.
- Segments are provided in context, allowing for document-level evaluation.

3.2 Creation of Reference Translations

The data acquisition process was structured into three steps to ensure high quality, consistency, and adherence to idiom-specific conventions.

1. **Translation**: We hired language professionals who are native speakers of both German and the respective Romansh idiom.
2. **Review**: Two expert linguists of Lia Rumantscha reviewed a sample of translations for a representative selection of varieties, and formulated feedback that was communicated to all translators.

↓ pred	gold → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	764	83	23	43	29	35
Surs.	98	810	67	46	55	54
Suts.	8	12	809	20	6	10
Surm.	19	9	32	811	8	6
Puter	10	12	12	17	648	30
Vall.	61	34	17	23	214	825

Table 2: Confusion matrix of a Romansh language variety classifier when applied to the reference translations.

↓ sys	ref → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG		60.0	47.5	53.8	47.2	49.4
Surs.	60.7		54.5	49.5	43.1	43.0
Suts.	48.3	54.7		50.9	39.5	39.1
Surm.	54.8	49.8	51.0		43.1	43.3
Puter	47.4	42.7	39.0	42.4		58.7
Vall.	49.3	42.4	38.4	42.4	58.4	

Table 3: Pairwise ChrF scores between the reference translations for the different varieties.

3. **Revision**: The translators incorporated the feedback into the reference translations.

We provided the translators with a guidelines document, inspired by the WMT24 translator brief (Kocmi et al., 2024). The key points of the guidelines, which we provide in Appendix D, are:

- The German text is the main source for the translation into Romansh, while the English text can be used as an additional reference in case of ambiguity.
- No AI tools should be used for the translation.

The translators and reviewers had access to the complete context of each segment, including a link to the original website from which the segment was extracted (e.g., for segments from the Speech domain, the original YouTube video).

3.3 Challenges in the Data Acquisition

A challenge we encountered in the translation process was that the degree of standardization can vary across text domains. The Romansh idioms are well-standardized, which is reflected in the formal domains *News* and (partially) *Literary*. However, in the *Social* and *Speech* domains, there is more room for individual variation based on the translator's style or regional background. Therefore, while we consider the reference translations suitable for their intended use of evaluating idiom-aware MT, the dataset does not aim to represent the full spectrum of variation present in the Romansh idioms.

4 Validation Experiments

We perform two automatic validation experiments to confirm that the reference translations are suitable for variety-specific evaluation:

Language Classification We use a fastText classifier (Joulin et al., 2017) trained on a corpus of Romansh newspaper articles that were manually labeled with their variety. Table 2 shows that when applied to our reference translations, the classifier predicts the expected variety for the majority of segments, indicating that the reference translations exhibit variety-specific features.

Cross-Variety Scores We calculate pairwise ChrF scores (Popović, 2015) between the sets of references, which are reported in Table 3. The maximum ChrF score across varieties is 60.7 (for Sursilvan–Rumantsch Grischun), which confirms that the sets of reference translations are distinct from each other even for related varieties, allowing for variety-specific evaluation. At the same time, the cross-variety scores are high enough to rule out serious data quality issues, such as a systematic misalignment of segments.

5 Evaluation of MT Systems and LLMs on Translation from and into Romansh

We use our benchmark to evaluate the following machine translation systems and LLMs:

- **MADLAD-400** (Kudugunta et al., 2023), a family of open-source MT models trained on parallel data in more than 450 languages, including Romansh. We report results for the largest, 10.7B-parameter model, using sentence segmentation with SpaCy to translate sentences individually, with a beam size of 5.
- **Supertext**, a commercial MT system that supports German and Romansh, among other languages.[3] We use the website of Supertext to translate the segments in an Excel file.
- **Translatur-ia**, a closed, early prototype of an MT system that translates from German into

[3] https://supertext.com/

System	Rumantsch Grischun	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader
MADLAD-400 (10.7B)						
– direct	58.3 / 63.0	52.9 / 54.7	40.6 / 38.1	45.2 / 40.4	49.7 / 49.8	52.8 / 52.9
– pivoting via English	56.1 / 64.9	50.3 / 52.8	39.4 / 37.4	42.4 / 40.1	47.2 / 49.8	49.3 / 51.9
Supertext	72.3 / 92.6	66.9 / 90.7	58.7 / 76.6	62.9 / 81.5	67.0 / 85.2	69.1 / 86.6
Llama 3.3 (70B)	63.1 / 82.8	57.0 / 75.5	48.7 / 59.2	52.1 / 64.3	57.1 / 73.2	60.0 / 75.4
GPT-4o	74.3 / 92.9	70.9 / 92.2	64.2 / 85.2	67.7 / 87.3	71.7 / 90.6	75.1 / 91.1
Gemini 2.5 Flash	**75.4 / 93.1**	**72.1 / 92.9**	**68.5 / 89.4**	**71.7 / 90.6**	**73.5 / 91.7**	**77.7 / 92.3**

Table 4: **Romansh as source language:** ChrF / xCOMET scores of MT systems and LLMs for translation into German from six varieties of Romansh.

System	Rumantsch Grischun	Sursilvan	Sutsilvan	Surmiran	Puter	Vallader
MADLAD-400 (10.7B)						
– direct	48.0	40.7	34.6	37.1	37.0	38.5
– pivoting via English	50.7	43.0	36.1	38.7	38.6	40.1
Translatur-ia	19.7	18.1	16.7	17.4	17.3	17.6
Supertext	**68.9**	53.2	43.5	47.8	46.7	49.0
Llama 3.3 (70B)	52.1	43.9	36.6	39.3	40.3	42.6
GPT-4o	64.8	**60.1**	41.4	46.4	52.3	55.9
Gemini 2.5 Flash	66.0	58.7	**43.7**	**50.1**	**53.8**	**57.2**

Table 5: **Romansh as target language:** ChrF scores of MT systems and LLMs for translation from German into Romansh. Results in gray are based on translations into Rumantsch Grischun, which is the only target variety officially supported by these systems.

Rumantsch Grischun.[4]

- **Llama 3.3** (Grattafiori et al., 2024), an open-source LLM released in November 2024. We use the 70B-parameter version.
- **GPT-4o** (OpenAI et al., 2024), a commercial LLM that was released in May 2024 and was billed at $2.50 per million input tokens and $10 per million output tokens.
- **Gemini 2.5 Flash** (Comanici et al., 2025), a commercial LLM that was released in June 2025 and was billed at $0.30 per million input tokens and $2.50 per million output tokens. We turn off the 'thinking' mode to enable a direct comparison with the other systems.

LLM Prompting When using LLMs for translation, we use the same prompting setup as the WMT24 General Machine Translation Shared Task (Kocmi et al., 2024).[5] Specifically, we use 3-shot prompting with temperature set to zero. The prompt template is listed in Appendix B.[6] As few-shot examples, we use typical example sentences from the fable *The Fox and the Crow* (Gross, 2004), which we list in Appendix C.

Quality Metrics For evaluating translation quality, we use ChrF (Popović, 2015), a metric based on character n-grams that does not require word segmentation, via SacreBLEU (Post, 2018).[7]

For evaluating translations from Romansh into German, we additionally use xCOMET (Guerreiro et al., 2024), a neural metric that was ranked highly in the WMT24 Metrics Shared Task (Freitag et al., 2024). We use model version `XCOMET-XL`[8] in the *reference-only* mode, i.e., we do not provide the Romansh source sequence to the metric, a language it

[4] https://translaturia.fhgr.ch/

[5] https://github.com/wmt-conference/wmt-collect-translations

[6] A limitation of this prompt template is that it does not provide the LLM with context beyond the segment that is currently being translated. We opt to keep the setup similar to WMT24 and leave document-level evaluation to future work.

[7] Signature: `#:1|c:mixed|e:yes|nc:6|nw:0|s:no|v:2.5.1`

[8] https://hf.co/Unbabel/XCOMET-XL

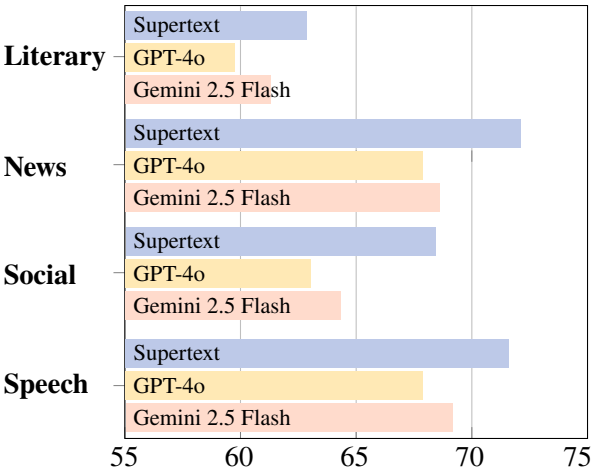


Figure 2: Domain-specific ChrF scores of systems translating from German into Rumantsch Grischun.

↓ **tgt**	**ref** → RG	Surs.	Suts.	Surm.	Puter	Vall.
RG	66.0	51.6	42.6	46.7	46.2	48.5
Surs.	57.0	58.7	42.6	44.8	43.7	45.1
Suts.	56.7	50.3	43.7	46.6	44.1	45.8
Surm.	52.3	46.5	42.5	50.1	43.8	44.5
Puter	47.5	42.1	38.0	41.2	53.8	54.6
Vall.	49.9	43.2	38.6	41.9	52.7	57.2

Figure 3: "Confusion matrix" of Gemini 2.5 Flash when translating into specific Romansh varieties. To visualize the degree to which the LLM output matches the requested variety (**tgt**), we evaluate the outputs with each set of reference translations (**ref**). A system that adheres to the requested target variety will achieve higher ChrF scores in the diagonal cells than in the off-diagonal cells.

has not seen during training. While xCOMET is designed to support such monolingual, *reference-only* evaluation, this mode has not been as extensively validated as other modes. Thus, xCOMET complements ChrF but should be interpreted with some caution. Following Kocmi et al. (2024), we report xCOMET as the macro-average over domains to control for different segment granularities.

6 Results

6.1 Comparison of Translation Directions

Table 4 shows the results for translation from Romansh into German, while Table 5 shows the results for translation from German into Romansh. The former performs consistently better than the latter; this is observed both for the supervised MT systems (MADLAD, Supertext) and for the LLMs.

Comparing performance across the six varieties of Romansh, we find that translation out of Romansh into German is relatively robust to linguistic variation: For Gemini, the gap in terms of ChrF between the minimum and maximum is $77.7 - 68.5$. In contrast, for translation into Romansh, the gap is $66.0 - 43.7$. Future work could exploit this asymmetry by using back-translation (Sennrich et al., 2016) for augmenting monolingual Romansh text with synthetic German translations.

6.2 Ranking of Models

For translation from Romansh into German, we report both ChrF and xCOMET scores in Table 4. We find that the system rankings are largely consistent between the two metrics, on average over the four domains, with Gemini 2.5 Flash achieving the highest scores according to both metrics. MADLAD-400 underperforms the other systems, likely due to the limited Romansh training data and the massively multilingual nature of the model.

In the German–Romansh direction, where ChrF is the only available metric (Table 5), we find that Gemini 2.5 Flash again achieves the highest scores for four out of six varieties. Supertext is the highest-ranked system for translation into Rumantsch Grischun, which is the officially supported target variety of this product.

6.3 Domain Difficulty

Figure 2 compares ChrF scores for German–Rumantsch Grischun translation for the four domains covered by the WMT24++ benchmark. The figure indicates that the *News* domain is the least challenging for all systems, which is consistent with findings of the WMT24 task for other languages (Kocmi et al., 2024). Surprisingly, lowest ChrF scores are achieved in the *Literary* domain, while the human evaluation of the WMT24 task did not find a systematic difference between the *Literary* and *News* domains in terms of difficulty. The *Speech* domain yields similar scores to *News*, and *Social* is slightly more challenging. Detailed results for each domain and variety (Appendices E and F) indicate that this pattern is consistent across varieties.

6.4 Target Variety Adherence of LLMs

While MADLAD and Supertext are limited to Rumantsch Grischun as the target variety, the LLMs can be prompted to produce translations in any of the six varieties. This raises the question of

whether the LLMs actually adhere to the requested target variety. Figure 3 shows a "confusion matrix" for Gemini 2.5 Flash, where we evaluate the system output not only with the reference translations for the requested target variety, but also with contrastive reference translations for the other varieties. The results suggest that state-of-the-art LLMs already have some degree of idiom awareness, but gravitate towards the higher-resource varieties (Rumantsch Grischun, Sursilvan, and Vallader).

7 Conclusion

This work fills a long-standing gap in the evaluation of machine translation for the Romansh language: the creation of a benchmark for the six main varieties of Romansh, and the provision of baseline results for existing MT systems and LLMs that cover Romansh.

Acknowledgements

We thank RTR and Fundaziun Patrimoni Cultural RTR for their support. We are grateful to Kirill Semenov for assistance with the poster presentation, Zachary Hopton for assistance with editing the manuscript, Noëmi Aepli and Martin Volk for helpful advice, and Tom Kocmi for help with the WMT24 benchmark.

Figure 1 uses maps from the Canton of Grisons and from Wikimedia Commons (User:Tomchen1989, User:NordNordWest, User:TUBS, CC BY-SA 3.0).

Author Contributions

JV: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft.
IPP: Funding acquisition, Project administration, Supervision, Writing – review & editing.
NBS: Writing – original draft.
SBG, AB, SB, MC, GPG, FH, GH, AL, VL, WR: Translation and/or translation quality assurance.
LD, BV: Linguistic supervision of translation workflow.
AR: Software.
RS: Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

References

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3284 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Renata Coray. 2008. *Von der Mumma Romontscha zum Retortenbaby Rumantsch Grischun : rätoromanische Sprachmythen*. Cultura alpina. Institut für Kulturforschung Graubünden, Chur.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Hans Goebl. 2003. Externe Sprachgeschichte der romanischen Sprachen im Zentral-und Ostalpenraum. *Romanische Sprachgeschichte: Ein internationales Handbuch zur Geschichte der romanischen Sprachen*, 1:747–773.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh

Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Manfred Gross. 2004. *Romansh: Facts & Figures*, [2nd rev. and updated ed.] edition. Lia rumantscha, Chur.

Matthias Grünert. 2018. Multilingualism in Switzerland. In *Manual of Romance Sociolinguistics*, pages 526–548. De Gruyter. Section: Manual of Romance Sociolinguistics.

Matthias Grünert. 2024. Rätoromanisch. In Elvira Glaser, Johannes Kabatek, and Barbara Sonnenhauser, editors, *Sprachenräume der Schweiz. Band 1: Sprachen*, pages 156–184. Narr Francke Attempto, Tübingen.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

John Haiman and Paola Benincà. 1992. *The rhaetoromance languages*. Routledge, London.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. glottolog/glottolog: Glottolog database 5.2.1.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Ricarda Liver. 2010. *Rätoromanisch: eine Einführung in das Bündnerromanische*, 2., überarbeitete und erweiterte Auflage edition. Narr Studienbücher. Narr Verlag, Tübingen.

Ricarda Liver. 2014. Le romanche des Grisons. In *Manuel des langues romanes*, pages 413–446. De Gruyter. Section: Manuel des langues romanes.

Christopher Moseley and Alexandre Nicolas. 2010. *Atlas of the world's languages in danger*, 3rd ed., entirely revised, enlarged and updated edition. UNESCO, Paris. Series: Memory of peoples series Book Title: Atlas of the world's languages in danger.

Žarko Muljačić. 2012. Über den Begriff Dachsprache. In Ulrich Ammon, editor, *Status and Function of Languages and Language Varieties*, pages 256–277. De Gruyter.

Mathias Müller, Annette Rios, and Rico Sennrich. 2020. Domain robustness in neural machine translation. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 151–164, Virtual. Association for Machine Translation in the Americas.

Joel Niklaus, Jakob Merane, Luka Nenadic, Sina Ahmadi, Yingqiang Gao, Cyrill A. H. Chevalley, Claude Humbel, Christophe Gösken, Lorenzo Tanzi, Thomas Lüthi, Stefan Palombo, Spencer Poff, Boling Yang, Nan Wu, Matthew Guillod, Robin Mamié, Daniel Brunner, Julio Pereyra, and Niko Grupen. 2025. SwiLTra-bench: The Swiss legal translation benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14894–14916, Vienna, Austria. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Philipp Obrist. 2022. Discourse traditions in the history of Romansh. In *Manual of Discourse Traditions in Romance*, pages 615–632. De Gruyter. Section: Manual of Discourse Traditions in Romance.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, and 400 others. 2024. GPT-4o system card. *Preprint*, arXiv:2410.21276.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on*

Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Yves Scherrer and Bruno Cartoni. 2012. The trilingual ALLEGRA corpus: Presentation and possible use for lexicon induction. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2890–2896, Istanbul, Turkey. European Language Resources Association (ELRA).

Heinrich Schmid. 1976. Zur Gliederung des Bündnerromanischen. *Annalas da la Societad Retorumantscha*, 89:7–62. Publisher: Stampa Romontscha.

Heinrich Schmid. 1982. *Richtlinien für die Gestaltung einer gesamtbündnerromanischen Schriftsprache: Rumantsch grischun*, [2. aufl.] edition. Lia Rumantscha, Cuira.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

A Dataset Statistics

Variety	Segments					Tokens				
	Lit.	**News**	**Soc.**	**Speech**	**Total**	**Lit.**	**News**	**Soc.**	**Speech**	**Total**
German (Deutsch et al., 2025)	206	149	531	111	998	9 814	9 416	10 556	10 157	39 954
RG	206	149	531	111	998	10 893	11 626	11 949	10 174	44 653
Sursilvan	206	149	531	111	998	10 609	11 287	11 913	10 144	43 964
Sutsilvan	206	149	531	111	998	11 147	11 667	12 241	10 378	45 444
Surmiran	206	149	531	111	998	10 718	11 424	11 907	10 167	44 227
Puter	206	149	531	111	998	11 325	11 595	12 267	10 490	45 688
Vallader	206	149	531	111	998	11 412	11 700	12 330	10 513	45 966

Table 6: Dataset statistics for each language variety. Token counts are based on whitespace tokenization.

B Prompt Template for LLMs

The below example is parameterized as follows:

- Source language: German
- Target language: Romansh
- Target variety: Rumantsch Grischun
- Source sentence: *"Sisos Darstellungen von Land und Wasser in neuer Ausstellung"*

````
Translate the following segment surrounded in triple backticks into Romansh
(Rumantsch Grischun variety). The German segment:

```Sisos Darstellungen von Land und Wasser in neuer Ausstellung```
````
````

C Few-shot Examples for LLM Prompting

German:

Der Fuchs war wieder einmal hungrig.

Da sah er auf einer Tanne einen Raben, der ein Stück Käse in seinem Schnabel hielt.

Das würde mir schmecken, dachte er, und rief dem Raben zu: «Wie schön du bist! Wenn dein Gesang ebenso schön ist wie dein Aussehen, dann bist du der Schönste von allen Vögeln».

English:

The fox was hungry yet again.

There, he saw a raven upon a fir holding a piece of cheese in its beak.

This I would like, he thought, and shouted at the raven: "You are so beautiful! If your singing is as beautiful as your looks, then you are the most beautiful of all birds."

Rumantsch Grischun:

La vulp era puspè ina giada fomentada.

Qua ha ella vis sin in pign in corv che tegneva in toc chaschiel en ses pichel.

Quai ma gustass, ha ella pensà, ed ha clamà al corv: «Tge bel che ti es! Sche tes chant è uschè bel sco tia parita, lur es ti il pli bel utschè da tuts».

Sursilvan:

L'uolp era puspei inagada fomentada.

Cheu ha ella viu sin in pegn in tgaper che teneva in toc caschiel en siu bec.

Quei gustass a mi, ha ella tertgau, ed ha clamau al tgaper: «Tgei bi che ti eis! Sche tiu cant ei aschi bials sco tia cumparsa, lu eis ti il pli bi utschi da tuts».

Sutsilvan:

La gualp eara puspe egn'eada fumantada.

Qua â ella vieu sen egn pegn egn corv ca taneva egn toc caschiel ainten sieus pecel.

Quegl gustass a mei, â ella tartgieu, ed ha clamo agli corv: «Tge beal ca tei es! Scha tieus tgànt e aschi beal sco tia pareta, alura es tei igl ple beal utschi da tuts».

Surmiran:

La golp era puspe eneda famantada.

Co ò ella via sen en pegn en corv tgi tigniva en toc caschiel an sies pecal.

Chegl am gustess, ò ella panso, ed ò clamo agl corv: «Tge bel tgi te ist! Schi ties cant è schi bel scu tia parentscha, alloura ist te igl pi bel utschel da tots».

Puter:

La vuolp d'eira darcho üna vouta famanteda.

Co ho'la vis sün ün pin ün corv chi tgnaiva ün töch chaschöl in sieu pical.

Que am gustess, ho'la penso, ed ho clamo al corv: «Che bel cha tü est! Scha tieu chaunt es uschè bel scu tia apparentscha, alura est tü il pü bel utschè da tuots».

Vallader:

La vuolp d'eira darcheu üna jada fomantada.

Qua ha'la vis sün ün pin ün corv chi tgnaiva ün toc chaschöl in seis pical.

Quai am gustess, ha'la pensà, ed ha clomà al corv: «Che bel cha tü est! Scha teis chant es uschè bel sco tia apparentscha, lura est tü il plü bel utschè da tuots».

D Translation Guidelines (in German)

1. Wofür werden die Übersetzungen benötigt?

Wir verwenden die von Ihnen erstellen Übersetzungen für die Evaluierung von maschinellen Übersetzungssystemen. Der Output der Übersetzungssysteme wird mit Ihrer Übersetzung verglichen – je ähnlicher der Output, desto besser das Übersetzungssystem.

Die Referenzübersetzungen werden nicht für das Training der Übersetzungssysteme verwendet.

2. Wie wurden die Texte ausgewählt?

Die Texte stammen aus einem bestehenden Datensatz («WMT24»), der zuvor bereits aus dem Englischen in 55 verschiedene Sprachen übersetzt worden ist, darunter Deutsch.

Der Datensatz setzt sich aus vier Textsorten zusammen:

- **literary:** Fan-Fiction, welche auf der Website «Archive of Our Own» veröffentlicht wurde.
- **news:** Zufällig ausgewählte Online-News vom Januar 2024.
- **social:** Zufällig ausgewählte Threads aus dem Sozialen Netzwerk «Mastodon».
- **speech:** Transkripte zufällig ausgewählter YouTube-Videos.

3. Wie ist die Excel-Datei aufgebaut?

Wir erstellen für jedes Idiom eine eigene Excel-Datei. Die Datei enthält vier Tabellen für die vier Textsorten, und jede Zeile in der Tabelle entspricht einem Textsegment. Mehrere Textsegmente setzen sich zu zusammenhängenden Dokumenten zusammen, welche durch graue Leerzeilen voneinander abgetrennt sind.

Die Tabellen sind wie folgt aufgebaut:

- **English:** Originaler Text auf Englisch.
- **document_id:** ID des Dokuments.
- **segment_id:** ID des Segments.
- **url:** Webseite, von welcher das Dokument ursprünglich bezogen wurde. Kann optional aufgerufen werden, um den Kontext des Dokuments nachzuvollziehen.
- **German:** Referenzübersetzung auf Deutsch.
- **translation:** Hier soll die Übersetzung auf Rätoromanisch eingetragen werden.
- **comment:** Kann von Ihnen optional benutzt werden, um uns einen wichtigen Kommentar zu hinterlassen.

4. Welche Anforderungen gelten an die Übersetzungen?

- Die Referenzübersetzungen repräsentieren Ihre Erwartungen an den Output eines guten maschinellen Übersetzungssystems. Überlegen Sie sich: Was für eine Übersetzung würde ein gutes Übersetzungssystem (im Stil von DeepL oder Google Translate) für das entsprechende romanische Idiom erzeugen?
- Massgeblich für Ihre Übersetzung sollte der Text auf Deutsch sein. Der englische Text kann optional zu Rate gezogen werden, falls der deutsche Text mehrdeutig ist.
- Weil wir Ihre Referenzübersetzungen für die Evaluierung verwenden möchten, bitten wir Sie, keine AI-Tools zu verwenden. Bitte erstellen Sie die Übersetzungen von Grund auf in Ihren eigenen Worten.

  - Erlaubte Hilfsmittel: Wörterbücher, Translation Memories, ...
  - Nicht erlaubte Hilfsmittel: ChatGPT, Copilot, Supertext/Textshuttle, DeepL, ...

  Wir können den Übersetzungen ansehen, wenn sie mit AI-Unterstützung erstellt wurden, und wir müssten im schlimmsten Fall die Übersetzungen noch einmal neu erstellen lassen (dies ist leider in der Vergangenheit schon vorgekommen).

- Bitte fügen Sie keine Übersetzungsalternativen oder Erklärungen in Klammern in die Übersetzungen ein.
- Bitte übersetzen Sie die Information in den Texten vollständig.
- Lehnwörter, Eigennamen etc. dürfen gerne auf Englisch oder Deutsch belassen werden, falls ein gutes rätoromanisches Übersetzungssystem das gleiche machen sollte.
- Betreffend die Textsorte «social»:
  - Viele Posts enthalten Fachbegriffe, Slang, Abkürzungen. Durch die Übersetzung vom Englischen ins Deutsche dürfte vieles schon weniger kryptisch geworden sein. Falls dennoch bei einem Post Unsicherheit besteht, können Ihnen diese Notizen weiterhelfen: https://github.com/wmt-conference/wmt24-news-systems/blob/main/README-social-domain-translation-notes.pdf
  - Nutzernamen wurden anonymisiert (z.B. @user1, @user2). Bitte übernehmen Sie die Nutzernamen eins zu eins, d.h. Sie müssen diese nicht übersetzen.
  - Hingegen dürfen #Hashtags gerne übersetzt werden, falls dies im Kontext Sinn macht.

E Detailed Results for RM–DE

E.1 Rumantsch Grischun to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	50.6 / 52.8	67.9 / 81.3	53.0 / 71.7	59.5 / 46.4	57.8 / 63.0
– pivoting via English	49.8 / 57.1	63.2 / 79.1	51.9 / 73.1	58.0 / 50.2	55.7 / 64.9
Supertext	68.3 / 89.9	71.1 / 95.1	73.9 / **96.2**	75.8 / 89.1	72.3 / 92.6
Llama 3.3 (70B)	56.8 / 75.3	68.0 / 91.3	58.2 / 84.3	68.6 / 80.4	62.9 / 82.8
GPT-4o	**71.5** / **91.8**	74.1 / **95.9**	73.6 / 94.3	78.3 / 89.6	74.4 / 92.9
Gemini 2.5 Flash	**71.5** / 90.6	**74.4** / 95.7	**76.5** / 95.9	**79.5** / **90.2**	**75.5** / **93.1**

Table 7: ChrF / xCOMET scores for translation from Rumantsch Grischun into German.

E.2 Sursilvan to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	45.7 / 44.0	64.4 / 76.6	46.9 / 65.8	52.2 / 32.5	52.3 / 54.7
– pivoting via English	44.0 / 45.0	59.1 / 70.2	45.2 / 65.8	51.1 / 30.4	49.9 / 52.8
Supertext	63.6 / 87.1	68.7 / 94.8	64.5 / 94.4	70.6 / 86.4	66.9 / 90.7
Llama 3.3 (70B)	50.5 / 63.7	64.3 / 88.4	50.4 / 77.0	61.7 / 72.9	56.7 / 75.5
GPT-4o	68.6 / **90.7**	72.2 / 95.5	67.8 / 93.0	**74.9** / **89.5**	70.9 / 92.2
Gemini 2.5 Flash	**69.7** / 90.5	**72.9** / **95.7**	**70.8** / **95.7**	**74.9** / **89.5**	**72.1** / **92.9**

Table 8: ChrF / xCOMET scores for translation from Sursilvan into German.

E.3 Sutsilvan to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	33.1 / 31.4	54.1 / 50.3	32.8 / 48.4	40.0 / 22.3	40.0 / 38.1
– pivoting via English	32.9 / 31.5	50.3 / 47.2	33.1 / 49.8	39.4 / 21.2	38.9 / 37.4
Supertext	55.1 / 67.9	62.9 / 87.8	54.7 / 86.1	61.1 / 64.8	58.5 / 76.6
Llama 3.3 (70B)	43.0 / 44.3	57.6 / 74.8	40.3 / 61.8	52.9 / 55.7	48.4 / 59.2
GPT-4o	61.5 / 80.5	70.0 / 93.3	54.4 / 81.6	70.3 / 85.2	64.1 / 85.2
Gemini 2.5 Flash	**65.7** / **85.0**	**71.2** / **94.2**	**64.6** / **92.4**	**72.3** / **86.1**	**68.5** / **89.4**

Table 9: ChrF / xCOMET scores for translation from Sutsilvan into German.

E.4 Surmiran to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	37.8 / 34.1	58.6 / 53.0	37.9 / 52.7	43.9 / 21.7	44.5 / 40.4
– pivoting via English	36.9 / 36.1	53.0 / 50.4	36.0 / 50.6	42.0 / 23.1	41.9 / 40.1
Supertext	59.6 / 76.9	67.8 / 91.6	58.9 / 88.6	64.4 / 69.0	62.7 / 81.5
Llama 3.3 (70B)	46.7 / 52.9	60.9 / 78.7	44.2 / 66.0	55.2 / 59.7	51.7 / 64.3
GPT-4o	66.4 / 86.2	71.6 / 92.3	61.3 / 85.6	71.5 / 85.1	67.7 / 87.3
Gemini 2.5 Flash	**69.3 / 88.5**	**72.8 / 94.3**	**69.6 / 93.0**	**75.2 / 86.7**	**71.7 / 90.6**

Table 10: ChrF / xCOMET scores for translation from Surmiran into German.

E.5 Puter to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	42.6 / 41.0	63.3 / 68.7	41.8 / 58.6	48.2 / 30.9	49.0 / 49.8
– pivoting via English	41.9 / 42.9	57.1 / 63.2	40.4 / 60.2	47.4 / 32.9	46.7 / 49.8
Supertext	63.9 / 81.6	71.3 / 93.4	62.6 / 89.4	69.5 / 76.4	66.8 / 85.2
Llama 3.3 (70B)	52.5 / 63.6	64.6 / 84.8	49.0 / 73.8	61.0 / 70.4	56.8 / 73.2
GPT-4o	69.7 / 88.7	73.6 / 94.0	67.0 / 91.7	76.6 / **88.0**	71.7 / 90.6
Gemini 2.5 Flash	**71.5 / 89.7**	**74.6 / 94.6**	**70.7 / 94.5**	**77.1 / 88.0**	**73.5 / 91.7**

Table 11: ChrF / xCOMET scores for translation from Puter into German.

E.6 Vallader to German

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	45.9 / 46.9	65.0 / 71.5	46.1 / 63.8	51.7 / 29.3	52.2 / 52.9
– pivoting via English	43.1 / 47.8	59.6 / 65.7	43.7 / 64.3	48.7 / 29.7	48.8 / 51.9
Supertext	65.6 / 83.3	71.6 / 92.5	66.7 / 91.7	72.1 / 78.9	69.0 / 86.6
Llama 3.3 (70B)	54.7 / 67.2	67.3 / 86.2	52.8 / 78.6	64.1 / 69.5	59.8 / 75.4
GPT-4o	72.4 / 90.1	77.3 / 94.5	71.2 / 91.2	79.4 / **88.4**	75.1 / 91.1
Gemini 2.5 Flash	**74.7 / 91.1**	**78.9 / 94.7**	**76.1 / 95.3**	**80.9** / 88.1	**77.6 / 92.3**

Table 12: ChrF / xCOMET scores for translation from Vallader into German.

F Detailed Results for DE–RM

F.1 German to Rumantsch Grischun

System	Literary	News	Social	Speech	Macro-Average
MADLAD-400 10.7B					
– direct	42.9	56.3	42.6	48.6	47.6
– pivoting via English	45.6	56.6	47.6	51.8	50.4
Translatur-ia	17.1	20.1	23.4	17.2	19.5
Supertext	**62.8**	**72.1**	**68.4**	**71.6**	**68.7**
Llama 3.3 (70B)	46.8	57.7	48.6	54.0	51.8
GPT-4o	59.9	67.9	63.0	67.9	64.7
Gemini 2.5 Flash	61.3	68.6	64.3	69.2	65.8

Table 13: ChrF scores for translation from German into Rumantsch Grischun.

F.2 German to Sursilvan

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	39.7	50.7	39.4	44.5	43.6
GPT-4o	**56.2**	**63.7**	**57.5**	**62.4**	**59.9**
Gemini 2.5 Flash	55.4	61.8	56.4	60.6	58.5

Table 14: ChrF scores for translation from German into Sursilvan.

F.3 German to Sutsilvan

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	32.4	43.0	33.0	36.8	36.3
GPT-4o	37.1	47.5	39.6	40.1	41.1
Gemini 2.5 Flash	**40.5**	48.9	39.8	**44.5**	**43.4**

Table 15: ChrF scores for translation from German into Sutsilvan.

F.4 German to Surmiran

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	34.7	46.5	35.5	38.8	38.9
GPT-4o	42.2	53.5	43.1	44.8	45.9
Gemini 2.5 Flash	**46.9**	**55.7**	**47.0**	**49.6**	**49.8**

Table 16: ChrF scores for translation from German into Surmiran.

F.5 German to Puter

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	35.8	47.0	36.0	41.0	39.9
GPT-4o	48.0	57.5	**49.2**	53.4	52.0
Gemini 2.5 Flash	**51.3**	**59.9**	48.5	**54.8**	**53.6**

Table 17: ChrF scores for translation from German into Puter.

F.6 German to Vallader

System	Literary	News	Social	Speech	Macro-Average
Llama 3.3 (70B)	38.3	49.2	38.6	42.9	42.2
GPT-4o	52.7	60.2	53.2	56.5	55.7
Gemini 2.5 Flash	**53.8**	**61.6**	**54.2**	**58.2**	**57.0**

Table 18: ChrF scores for translation from German into Vallader.

G Examples for System Outputs

G.1 German to Romansh

[English:] *it seems like even iMessage over WiFi isn't working, which doesn't quite make sense to me*

German source: *Anscheinend funktioniert nicht mal iMessage über WiFi, was mir nicht ganz einleuchtet*

Rumantsch Grischun

Reference: *Para che gnanc iMessage funcziunia via WiFi, tge ch'è per mai betg dal tut evident.*

MADLAD-400 (10.7B) direct: *Sembla ch'iMessage na funcziunia betg sur WiFi, quai che n'ha nagin sens per mai*

Translatur-ia: *tuttinascha fitg lungascha betg pitschnas la kieni sur la qualafizistgeli , ma prognostitgà*

Supertext: *Para na funcziuna gnanc iMessage sur WiFi, quai che ma resplenda betg dal tut*

Llama 3.3 (70B): *Apparentamain funcziuna schiz ni iMessage sur WiFi, quai ch'ha a mai betg dal tut enclaus.*

GPT-4o: *Apparentamain na funcziuna gnanc iMessage sur WiFi, quai che na ma para betg del tut cler*

Gemini 2.5 Flash: *Apparentamain na funcziunescha gnanc iMessage sur WiFi, quai che na ma para betg dal tuttafatg cler.*

Sursilvan

Reference: *Sco ei para funcziunescha gnanc iMessage sur WiFi, quei ch'jeu sai buca propi capir.*

Llama 3.3 (70B): *Scheppa che sch'iva nean iMessage sur WiFi, quai che mi n'entuppa betg dal tut*

GPT-4o: *Apparentamein funcziunescha buc ne iMessage sur WiFi, quei cheu ei buc clar a mi*

Gemini 2.5 Flash: *Apparentamein funcziunescha gnanc iMessage sur WiFi, quei che jeu capeschel buca dal tut.*

Sutsilvan

Reference: *Para funcziunescha gnànc iMessage sur igl WiFi, tge ca fa betga propi sen tanor me*

Llama 3.3 (70B): *Apparentamain funcziuna schi betg iMessage sur WiFi, sche mi nu cumpreta betg del tut*

GPT-4o: *Apparentamein fonzionescha gnanc iMessage sur WiFi, tgei ca mei betg clarifica cumpleinamein.*

Gemini 2.5 Flash: *Apparentamagn na funcziunescha betg gnieu iMessage sur WiFi, tgei ca na clerescha betg dal tut a mei.*

Surmiran

Reference: *Scu para funcziunescha mianc iMessage sur WiFi, chegl tg'ia sa betg propi tgapeir*

Llama 3.3 (70B): *Apparentamain funcziuna betg ni ir iMessage sur WiFi, sche che m'èn betg del tut clar*

GPT-4o: *Apparentamaint funcziunescha gnanc iMessage sur WiFi, tge nun è dal tut cler per mai.*

Gemini 2.5 Flash: *Apparentamaintg na funcziunescha betg gniev iMessage sur WiFi, chegl tgi na m'è betg dal tuttafatg cler.*

Puter

Reference: *Pera cha nu funcziuna niauncha iMessage sur WiFi, che ch'eau nun incleg dal tuot*

Llama 3.3 (70B): *Apparentamain funcziuna betg era neir iMessage sur WiFi, cha m'ha betg dal tut en persvasiun*

GPT-4o: *A para cha iMessage nu funcziuna nemain sur WiFi, quai nu'm para fich cler*

Gemini 2.5 Flash: *Apparentamaing nu funcziunescha neir betg iMessage sur WiFi, che nu'm clerescha*

dal tuot.

Vallader

Reference: *Apparaintamaing nu funcziuna gnanca üna jada iMessage sur WiFi, quai chi nu'm voul propcha ir per testa*

Llama 3.3 (70B): *Apparentamain funcziunia neir il iMessage sur WiFi, quai ch'ha mì nua betg dal tut en la testa*

GPT-4o: *Apparentamaing nu funcziuna neir iMessage sur WiFi, quai nu'm para fich cler.*

Gemini 2.5 Flash: *Apparentamaing nu funcziunescha neir betg iMessage sur WiFi, che nu'm para dal tuot cler.*

G.2 Romansh to German

[English:] *it seems like even iMessage over WiFi isn't working, which doesn't quite make sense to me*

German reference: *Anscheinend funktioniert nicht mal iMessage über WiFi, was mir nicht ganz einleuchtet*

Rumantsch Grischun

Romansh source: *Para che gnanc iMessage funcziunia via WiFi, tge ch'è per mai betg dal tut evident.*

MADLAD-400 (10.7B) direct: *Damit iMessage über WLAN funktioniert, Ich weiß nicht, ob ich das jemals wiederholen werde.*

Supertext: *Anscheinend funktioniert iMessage nicht einmal über WiFi, was für mich nicht ganz selbstverständlich ist.*

Llama 3.3 (70B): *Dass nicht einmal iMessage über WiFi funktioniert, ist überhaupt nicht offensichtlich.*

GPT-4o: *Es scheint, dass nicht einmal iMessage über WiFi funktioniert, was für mich nicht ganz nachvollziehbar ist.*

Gemini 2.5 Flash: *Es scheint, dass nicht einmal iMessage über WLAN funktioniert, was für mich nicht ganz offensichtlich ist.*

Sursilvan

Romansh source: *Sco ei para funcziunescha gnanc iMessage sur WiFi, quei ch'jeu sai buca propi capir.*

MADLAD-400 (10.7B) direct: *Er stellt fest, dass die Wii-Fernbedienung nicht funktioniert, und fragt sich, was er damit anfangen soll.*

Supertext: *Anscheinend funktioniert iMessage über WiFi nicht einmal, was ich wirklich nicht verstehen kann.*

Llama 3.3 (70B): *Weil ich als Funktion habe, dass ich iMessage über WiFi nutzen kann, was ich aber nicht richtig verstehe.*

GPT-4o: *Anscheinend funktioniert iMessage nicht einmal über WiFi, was ich wirklich nicht verstehen kann.*

Gemini 2.5 Flash: *Anscheinend funktioniert iMessage nicht einmal über WLAN, was ich nicht wirklich verstehen kann.*

Sutsilvan

Romansh source: *Para funcziunescha gnànc iMessage sur igl WiFi, tge ca fa betga propi sen tanor me*

MADLAD-400 (10.7B) direct: *Die erste Singleauskopplung des Debütalbums, I Don't Want to Miss a Thing, erreichte Platz 1 der US-Charts.*

Supertext: *Anscheinend funktioniert nicht einmal iMessage über WiFi, was auf jeden Fall nicht*

funktioniert.

Llama 3.3 (70B): *Damit Funktionen wie iMessage über das WiFi funktionieren, was aber nicht gerade mein Ding ist*

GPT-4o: *Sogar iMessage funktioniert nicht über das WLAN, was für mich keinen Sinn ergibt.*

Gemini 2.5 Flash: *Anscheinend funktioniert iMessage nicht mal über WLAN, was für mich nicht wirklich Sinn macht*

Surmiran

Romansh source: *Scu para funcziunescha mianc iMessage sur WiFi, chegl tg'ia sa betg propi tgapeir*

MADLAD-400 (10.7B) direct: *2000: 1. Preis beim Wettbewerb für Jugendtheaterstücke der Sächsischen Staatsoper Unter den Linden 2001: 1. Preis beim Wettbewerb für Jugendtheaterstücke der Sächsischen Staatsoper.*

Supertext: *Anscheinend funktioniert iMessage über WiFi nicht einmal, was man wirklich nicht verstehen kann.*

Llama 3.3 (70B): *Um Funktionen wie iMessage über WiFi zu nutzen, muss man wissen, dass das nicht ganz umsonst ist*

GPT-4o: *Anscheinend funktioniert iMessage nicht über WiFi, das kann ich nicht wirklich verstehen.*

Gemini 2.5 Flash: *Anscheinend funktioniert mein iMessage nicht über WLAN, was ich nicht wirklich verstehen kann.*

Puter

Romansh source: *Pera cha nu funcziuna niauncha iMessage sur WiFi, che ch'eau nun incleg dal tuot*

MADLAD-400 (10.7B) direct: *Ich weiß nicht, ob iMessage über WLAN funktioniert, was ich nicht verstehe.*

Supertext: *Auch wenn iMessage über WiFi nicht funktioniert, ist das nicht ganz nachvollziehbar.*

Llama 3.3 (70B): *Denn nun funktioniert auch iMessage nicht mehr über WiFi, weil ich nicht mehr online bin*

GPT-4o: *Es scheint, dass iMessage nicht einmal über WiFi funktioniert, was ich überhaupt nicht verstehe.*

Gemini 2.5 Flash: *Aber dass iMessage auch über WLAN nicht funktioniert, das verstehe ich überhaupt nicht.*

Vallader

Romansh source: *Apparaintamaing nu funcziuna gnanca üna jada iMessage sur WiFi, quai chi nu'm voul propcha ir per testa*

MADLAD-400 (10.7B) direct: *Anscheinend funktioniert iMessage nicht einmal mehr über WLAN, was mich sehr ärgert.*

Supertext: *Anscheinend funktioniert iMessage über WiFi nicht einmal, was mir wirklich nicht einleuchtet.*

Llama 3.3 (70B): *Aber offensichtlich funktioniert nicht einmal iMessage über WiFi, was mich nicht weiter wundert*

GPT-4o: *Anscheinend funktioniert iMessage nicht einmal über WiFi, was mir wirklich nicht in den Kopf will.*

Gemini 2.5 Flash: *Anscheinend funktioniert iMessage über WLAN nicht einmal, was mir wirklich nicht in den Kopf will.*

A French Version of the OLDI Seed Corpus

Malik Marmonier **Benoît Sagot** **Rachel Bawden**
Inria, Paris, France
{firstname.lastname}@inria.fr

Abstract

We present the first French partition of the OLDI Seed Corpus, our submission to the WMT 2025 Open Language Data Initiative (OLDI) shared task. We detail its creation process, which involved using multiple machine translation systems and a custom-built interface for post-editing by qualified native speakers. We also highlight the unique translation challenges presented by the source data, which combines highly technical, encyclopedic terminology with the stylistic irregularities characteristic of user-generated content taken from Wikipedia. This French corpus is not an end in itself, but is intended as a crucial pivot resource to facilitate the collection of parallel corpora for the under-resourced regional languages of France.

1 Introduction

While state-of-the-art machine translation (MT) has made significant strides, progress has largely been concentrated on a handful of high-resource languages (Haddow et al., 2022). For many of the world's languages, development is hindered by a lack of reliable training data (Joshi et al., 2020). Techniques like backtranslation (Bertoldi and Federico, 2009; Bojar and Tamchyna, 2011; Sennrich et al., 2016) can help bridge this gap, but they typically require an initial high-quality parallel corpus to "kickstart" the process. To address this, the OLDI Seed Corpus was created (originally as the NLLB Seed Dataset) to provide a small but high-quality, professionally translated dataset for dozens of low-resource languages (NLLB Team et al., 2022; Maillard et al., 2023). The source material consists of approximately 6,000 English sentences sampled from a curated list of core Wikipedia articles, ensuring broad topic coverage (NLLB Team et al., 2022).

The WMT 2025 Open Language Data Initiative (OLDI) shared task builds directly on this effort, inviting the research community to expand these foundational open-source datasets to more languages. Our work answers this call by adding a French partition to the OLDI Seed Corpus. While French is a high-resource language, its selection as a pivot is strategic for our ultimate goal: facilitating the creation of parallel corpora for under-resourced regional languages of France (e.g., Francoprovençal, Occitan, Picard). Translators for these languages are overwhelmingly more likely to have a native command of French than of English. Most importantly, many of these languages exist in a diglossic relationship with French, which serves as the dominant language for most technical and formal domains. A French source text would, therefore, greatly simplify the complex terminological work inherent in translating the encyclopedic content found in the OLDI Seed Corpus, enabling the creation of direct calques and other word-formation strategies that are linguistically and culturally more congruent than those derived from English. With that in mind, we went to great lengths to rigorously verify the French technical terminology throughout the data creation process. By providing this carefully curated French corpus, we aim to establish a solid foundation for future translation efforts.

In addition to the final French partition, we also contribute a supplementary dataset containing the full set of translation hypotheses generated by the nine different MT systems and prompting techniques used in our workflow. This resource, which pairs multiple machine-generated outputs with a final, human-post-edited reference for each source segment, may be of particular use for research on preference optimization and quality estimation in the Wikipedia domain.[1]

We hope that our contribution will encourage further community-driven expansions of the OLDI Seed Corpus.

[1]We make this resource publicly available under CC BY-SA 4.0 license: https://github.com/mmarmonier/ACReFOSC

Proceedings of the Tenth Conference on Machine Translation, pages 1048–1060
November 8-9, 2025

2 Linguistic Overview

French is a Romance language that evolved from the Vulgar Latin spoken by the inhabitants of northern Gaul after the Roman conquest. Its development was significantly shaped by the Germanic invasions of the 5th century, particularly by the Franks, whose linguistic habits profoundly influenced the phonology and vocabulary of the emerging Gallo-Romance vernacular (Rickard, 2014). The earliest extant text, the Strasbourg Oaths, dates to 842. Over the subsequent centuries, the dialect of the Île-de-France region, known as Francien, gradually gained prestige due to the political and cultural centrality of Paris. This Parisian variety formed the basis of a standardized literary and administrative language that was progressively imposed throughout the kingdom, through a process legally enforced by the Ordinance of Villers-Cotterêts in 1539. This history of internal linguistic unification through political centralization later provided a model for its imposition as the language of administration and education throughout France's colonial empire.

Today, with over 321 million speakers, French is the fifth most spoken language in the world (OIF, 2022). However, this numerical strength, driven almost entirely by African demography, masks a more complex reality. Recent independent assessments point to a rebalancing, with French losing ground to English in several symbolic and high-value domains. In parts of West Africa, its constitutional standing has weakened, with countries like Mali and Burkina Faso downgrading it from an "official" to a "working" language in 2023–2024. In North Africa, educational policies increasingly favor English, while in major Anglophone countries, like the United Kingdom and the United States, learner numbers have seen long-term declines (Collen and Duff, 2025; Lusin et al., 2023). Furthermore, English overwhelmingly dominates high-prestige domains such as scientific publishing even within France's own research output (OST, 2024). While the language's global landscape is in flux, the variety used in this corpus, standard French as spoken in France, remains the most widely recognized norm for formal written communication.

3 Data Collection

The shared task guidelines for Seed data contributions permit the use of post-edited machine translation (MTPE). Given the high quality of modern MT systems for the English-French language pair, we adopted this workflow. A full professional translation from scratch was deemed prohibitively expensive. Counterintuitively perhaps for a "seed" corpus, the source text represents a non-trivial amount of written content; at 136,656 total words, its length is comparable to that of a novel, falling between Jane Austen's *Pride and Prejudice* and Charles Dickens' *A Tale of Two Cities*. We estimated that a professional translation would have required a substantial budget—around €50,000, not accounting for the added complexity of the varied and often highly technical terminology found in the source segments—that we felt would be more strategically allocated to future translation efforts into the low-resource regional languages of France. The MTPE approach therefore allowed us to produce a highly adequate French corpus while preserving resources for our long-term goals.

3.1 Machine translation

The source text was the English partition of the OLDI Seed Corpus. To generate a diverse set of initial translation hypotheses for post-editing, we made use of nine different MT systems and/or prompting techniques. Among these were four "traditional" sequence-to-sequence Transformer models (Vaswani et al., 2017). We used a standard bilingual OPUS-MT model (Tiedemann and Thottingal, 2020) trained on a large collection of open parallel corpora. We also used three larger multilingual models: two from the NLLB family, namely the 3.3B-parameter model and the 600M distilled version (NLLB Team et al., 2022), and the 3B-parameter model from the MADLAD-400 project (Kudugunta et al., 2023). For all four of these systems, translations were generated at the sentence level using beam search with a beam size of 4.

The remaining five translation hypotheses were generated using Large Language Models (LLMs).

Four hypotheses were generated using Llama 4 Scout, a recently released 109B-parameter Mixture-of-Experts model (Meta AI, 2025) with a remarkable 10M-token context window. The first and most straightforward approach was to translate each segment individually, though in a context-informed way. For this, we designed a detailed prompt that instructed the model to act as an expert translator of encyclopedic documents and provided it with a set of guidelines adapted from the official OLDI trans-

lation instructions.[2] A crucial aspect of the OLDI Seed corpus is that, while its segments are sourced from a limited number of Wikipedia articles, they are not necessarily contiguous. To account for this, the prompt supplied the model with preceding text from the source article as context (up to five segments), while explicitly stating that this context might not be directly adjacent to the segment being translated. We also prompted the model to use a chain-of-thought process before producing the final translation, which was to be enclosed in specific XML tags (<translation>...</translation>) for automatic retrieval (see Appendix A.1).

To take advantage of the model's large context window, the other three hypotheses from Llama 4 Scout were generated by translating at the document level. We reconstructed the source documents by grouping all segments from the OLDI Seed corpus that shared the same source URL in their metadata, ordering them by their numerical ID. This full-document text was then translated in a single prompt under three different conditions, with the model explicitly instructed to produce only the final translation without any preceding chain-of-thought. The first setting was identical to the segment-level approach, including the full set of translation guidelines (see Appendix A.2). The second was a contrastive ablation where we removed the OLDI guidelines from the prompt, in order to test the model's ability to follow a detailed translator's brief. For the third setting, we again used the full set of instructions, but also provided the model with the complete text of the corresponding French Wikipedia article as additional, in-domain context to inform its translation (see Appendix A.3).

The last translation hypothesis was produced by DeepSeek-R1 (DeepSeek-AI et al., 2025) via its web interface, at the document level, with prompts identical to the first document-level setting used with Llama 4 (guidelines, no corresponding French Wikipedia page; see Appendix A.2). Due to the model's safety filters, it refused to translate a handful of documents pertaining to religion, racism, or politics.

A final processing step was required for all document-level translations. The single block of translated text produced by the LLMs had to be segmented and re-aligned with the original source segments. While this process was largely automated, manual intervention was required for approximately 5% of the documents to correct errors typically resulting from skipped or hallucinated segments, or from the addition of extraneous newlines.

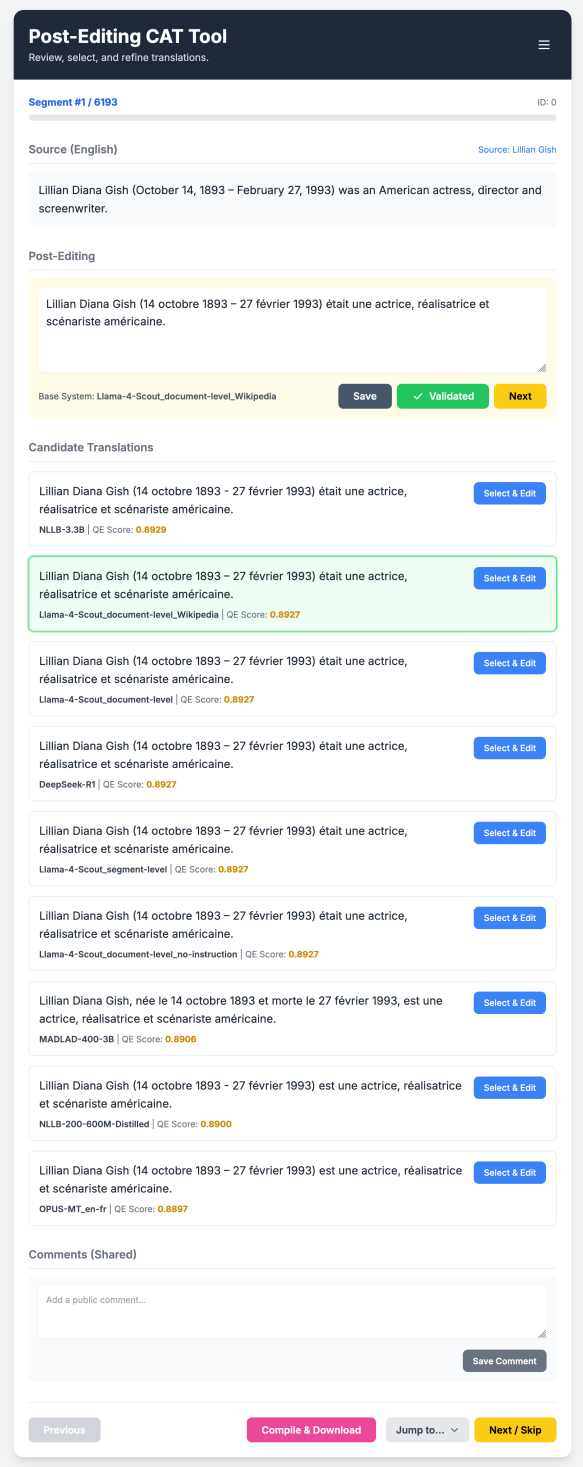


Figure 1: Our custom post-editing interface.

3.2 Human post-edition

The core of our contribution lies in a meticulous post-editing process, which was performed by two

[2] https://oldi.org/translation-guidelines.pdf

native French speakers with C2-level proficiency in English (the first and second authors). A native British English speaker with C2 proficiency in French (the third author) was also available for consultation to resolve ambiguities in the source text.

To facilitate this work, we developed a custom post-editing interface[3] (see Figure 1). For each source segment, the interface presents the user with all nine machine-generated hypotheses, sorted in descending order of their COMET[4] quality estimation (QE) scores (Rei et al., 2022). The post-editor can then select the most promising candidate, which populates a text area for refinement. This process allowed for an efficient workflow focused on two primary goals:

Fluency. Improving the naturalness and readability of the French text. A significant challenge throughout the post-editing process was dealing with various types of errors and disfluencies present in the English source text. These issues, likely stemming from the nature of user-generated content and the automated extraction process used to create the corpus, required careful interpretation and, at times, consultation with other language versions of the Seed corpus to resolve ambiguities. The problems ranged from simple typographical errors and ungrammatical constructions to more complex issues like garden-path sentences and segmentation errors that rendered segments nonsensical. Table 1 provides several examples of these challenges and our resulting post-edited translations.

Accuracy. Rigorously verifying and correcting the translation of technical terminology. Given the encyclopedic nature of the content, this required systematic external research. The corpus covers a dizzying array of technical topics, from cartography and bionanotechnology to Gothic architecture and Hilbert's problems, and verifying the correct terminology for each was a significant undertaking. Official resources like FranceTerme[5] were often of limited use. We therefore relied on extensive documentary research to find correct or acceptable equivalents in French, a crucial step to ensure the corpus can effectively serve as a pivot for the complex neology that will likely be required for translation into the regional languages of France. This task was made more difficult by the increasing prevalence of low-quality, machine-generated content on the web, or "AI slop," which degrades its utility as a reliable concordancer for the translator. In a handful of cases, we consulted with domain experts to resolve particularly challenging terminological issues. Table 2 provides examples of such challenging segments.

3.3 Human validation

After post-editing, all segments were processed through the Grammalecte[6] grammar checker interface (see Figure 2) for a final validation pass, correcting any residual spelling or grammatical errors.

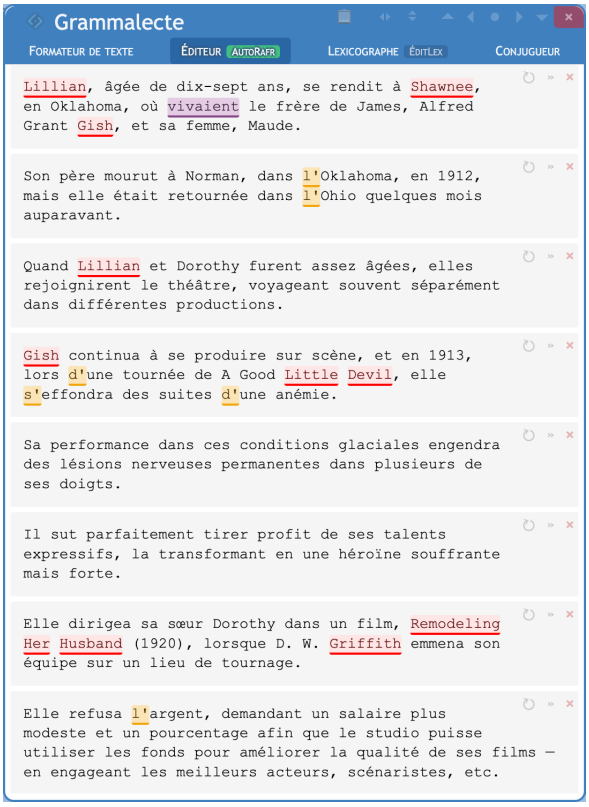


Figure 2: Grammalecte interface used for validation.

As per the shared task guidelines for Seed data, we have ensured that the terms of service for all MT models used allow for the reuse of their outputs. The final dataset is released under the same CC BY-SA 4.0 license as the source corpus.

4 Discussion

While we were initially tempted to follow the validation approach of Cols (2024) by training separate neural MT models on our final corpus and on each set of hypotheses, we ultimately judged that for a

[3]The interface was developed using Vue.js and styled with Tailwind CSS.

[4]wmt22-cometkiwi-da

[5]https://www.culture.fr/franceterme

[6]https://grammalecte.net/

ID	English Source & Issue Description
2129	**Source:** "Carleton University and the University of Western Ontario, 1945 and 1946 prospectively, created Journalism specific programs or schools." **Issue:** Typographical, lexical and grammatical errors ("prospectively" for "respectively", missing preposition, capitalization). **Post-edited:** "*L'Université Carleton et l'Université Western Ontario ont créé des programmes ou des écoles spécifiques de journalisme, respectivement en 1945 et 1946.*"
2244	**Source:** "Without social capital in the area of education, teachers and parents who play a responsibility in a students learning, the significant impacts on their child's academic learning can rely on these factors." **Issue:** Ungrammatical and confusing sentence structure (garden path, unclear pronoun reference). Resolved in part by soliciting the opinion of a native English speaker, and by following the interpretation of the Spanish version of the corpus. **Post-edited:** "*Sans capital social dans le domaine de l'éducation, et sans des enseignants et des parents qui jouent un rôle essentiel dans l'apprentissage de l'élève, les impacts significatifs sur l'apprentissage scolaire de leur enfant peuvent dépendre de ces facteurs.*"
3881	**Source:** "The first genetically modified ornamentals commercialized altered color." **Issue:** Garden-path sentence structure. Resolved by following the interpretation of the Italian version of the corpus. (We note that the Spanish version of the OLDI Seed has a different analysis of the segment.) **Post-edited:** "*Les premières plantes ornementales génétiquement modifiées commercialisées changeaient de couleur.*"
4326	**Source:** "When interest rates are very low, the number 0 is included if the interest rate is less than 1%, e.g. "% Treasury Stock", not "% Treasury Stock".)" **Issue:** Apparent error in the source text, likely from faulty extraction of a mathematical or financial example. The error was preserved as per the guidelines. **Post-edited:** "*Lorsque les taux d'intérêt sont très bas, le nombre 0 est inclus si le taux d'intérêt est inférieur à 1 %, par ex. « % Treasury Stock », et non « % Treasury Stock ».)*"
4539	**Source:** "Another early globe, the Hunt–Lenox Globe, ca." **Issue:** Sentence segmentation error; the segment is incomplete. **Post-edited:** "*Un autre des premiers globes, le globe Hunt–Lenox, env.*"

Table 1: Examples of challenges encountered in the source text during post-editing.

ID	English Source & Terminological Field
4044	**Source (Nanotechnology):** "Another group of nanotechnological techniques include those used for fabrication of nanotubes and nanowires, those used in semiconductor fabrication such as deep ultraviolet lithography, electron beam lithography, focused ion beam machining, nanoimprint lithography, atomic layer deposition, and molecular vapor deposition, and further including molecular self-assembly techniques such as those employing di-block copolymers." **Post-edited:** "*Un autre groupe de techniques nanotechnologiques comprend celles utilisées pour la fabrication de nanotubes et de nanofils, celles utilisées dans la fabrication de semi-conducteurs telles que la lithographie ultraviolette profonde, la lithographie par faisceau d'électrons, l'usinage par faisceau d'ions focalisés, la lithographie par nano-impression, le dépôt en couches atomiques et le dépôt moléculaire en phase vapeur, et incluant en outre des techniques d'auto-assemblage moléculaire telles que celles employant des copolymères à diblocs.*"
4845	**Source (Gothic Architecture):** "Lancet windows were supplanted by multiple lights separated by geometrical bar-tracery." **Post-edited:** "*Les fenêtres en lancette furent supplantées par des baies multiples séparées par des remplages géométriques.*"

Table 2: Examples of segments requiring specialized terminological research.

well-resourced language pair like English-French, and given that the initial MT hypotheses were already of a high standard—failing mostly on specific terminological choices and disfluencies inherited from the source—quality estimation using a state-of-the-art metric would provide a more telling as well as a more environmentally responsible validation of our data.

Although we had used a COMET-based QE model in our post-editing interface, we had observed certain limitations, such as insensitivity to terminological accuracy and a tendency to reward superficial trivial features (like hyphen or apostrophe type) at the expense of more essential features in the candidate translations. For our final validation, we therefore chose MetricX-24[7] (Juraska et al., 2024), a top-performing hybrid reference-based and reference-free metric. MetricX-24 is trained on human judgements (MQM and DA ratings; cf. Lommel et al. 2013; Graham et al. 2013) and predicts an error score, where lower scores indicate higher quality. This allows for a direct comparison of our final, human post-edited translations against the raw machine-generated hypotheses, thereby quantifying the value added by our human-in-the-loop process.

We scored our final post-edited translations and the raw outputs of the nine systems used to generate initial hypotheses. The results, including 95% confidence intervals and statistical significance groupings, are presented in Table 3.

As noted in Section 3.1, DeepSeek-R1 refused to translate a small number of documents due to its safety filters. To ensure a fair comparison of translation quality on the segments all systems were able to process, we also performed an analysis excluding these 165 segments. The results of this filtered evaluation are presented in Table 4.

Both analyses clearly validate the quality of our contributed dataset. The human post-edited text achieves the lowest average error score by a significant margin, placing it in a statistical group of its own (Group A), confirming the soundness and effectiveness of our manual post-editing and validation process. Among the machine-generated hypotheses, the results on the full dataset (Table 3) show a top tier (Group B) consisting of the larger sequence-to-sequence models and the segment-level Llama 4 Scout prompt. However, when excluding the segments DeepSeek-R1 refused to translate (Table 4), its relative performance improves significantly, moving it into this top tier of MT systems. This was not surprising, as we were consistently impressed during post-editing by DeepSeek-R1's knowledge of even the most arcane terminology, which we systematically verified through arduous external research ourselves. The remaining rankings are largely consistent across both analyses. Interestingly, all document-level prompting strategies for Llama 4 Scout were slightly less effective than the segment-level approach. The setting that ablated the specific OLDI guidelines ("no-instruction") performed on par with the standard document-level prompt from the point of view of MetricX-24; a direct comparison reveals that the specific OLDI guidelines had a limited impact on model generation, as these two settings produced identical translations in over 76% of cases, and for the minority of instances where they differed, the average Translation Edit Rate (TER; cf. Olive 2005; Snover et al. 2006) was a relatively low 9.48, indicating the variations were typically minor. Counter-intuitively, providing the model with the corresponding French Wikipedia article as additional context resulted in a statistically significant degradation in quality. Finally, the weakest-performing systems across all settings were the smaller distilled NLLB model and the bilingual OPUS-MT model.

Our post-editing logs offer another lens through which to evaluate the raw quality of the MT hypotheses. Of the 6,193 segments in the corpus, 3,043 (49.14%) had at least one machine-generated hypothesis that was deemed perfect by the post-editors and required no changes. A breakdown by system reveals that DeepSeek-R1 was by far the most reliable, providing the "perfect" translation in 2,503 instances, or 40.42% of all segments in the corpus. The other systems lagged considerably behind, with the various Llama 4 Scout prompting strategies and the larger NLLB and MADLAD models providing the perfect match in 7-9% of cases, while the smaller NLLB and OPUS-MT models did so less than 5% of the time.

5 Related Work

While data-driven approaches to MT have precedents as early as the 1950s (Edmundson and Hays, 1958), they did not truly come into their own until the late 1980s and early 1990s with the pioneering work on statistical machine translation (SMT) at

[7] metricx-24-hybrid-xl-v2p6

System	Avg. Error ↓	95% C.I.	Group
Human post-edition	2.0790	[2.04, 2.12]	A
NLLB-3.3B	2.2223	[2.19, 2.26]	B
MADLAD-400-3B	2.2290	[2.19, 2.27]	B
Llama-4-Scout_segment-level	2.2437	[2.21, 2.28]	B
Llama-4-Scout_document-level_no-instruction	2.3096	[2.27, 2.35]	C
Llama-4-Scout_document-level	2.3198	[2.28, 2.36]	C
Llama-4-Scout_document-level_Wikipedia	2.4322	[2.38, 2.48]	D
NLLB-200-600M-Distilled	2.5332	[2.49, 2.58]	E
DeepSeek-R1	2.5411	[2.48, 2.60]	E
OPUS-MT_en-fr	2.7019	[2.65, 2.75]	F

Table 3: MetricX-24 evaluation on the full dataset. Lower is better. Systems in the same lettered group are not statistically significantly different.

System	Avg. Error ↓	95% C.I.	Group
Human post-edition	2.0871	[2.05, 2.12]	A
DeepSeek-R1	2.2238	[2.18, 2.26]	B
NLLB-3.3B	2.2313	[2.19, 2.27]	B
MADLAD-400-3B	2.2386	[2.20, 2.28]	B
Llama-4-Scout_segment-level	2.2532	[2.22, 2.29]	B
Llama-4-Scout_document-level_no-instruction	2.3186	[2.28, 2.36]	C
Llama-4-Scout_document-level	2.3302	[2.29, 2.37]	C
Llama-4-Scout_document-level_Wikipedia	2.4456	[2.39, 2.50]	D
NLLB-200-600M-Distilled	2.5451	[2.50, 2.59]	D
OPUS-MT_en-fr	2.7099	[2.66, 2.76]	E

Table 4: MetricX-24 evaluation excluding segments refused by DeepSeek-R1. Lower is better. Systems in the same lettered group are not statistically significantly different.

IBM (Brown et al. 1990; Berger et al. 1994). This early research was heavily reliant on the availability of large parallel corpora, and the French-English language pair played a central role, largely thanks to the Canadian Hansard, a bilingual record of parliamentary proceedings. So influential was this corpus that it established a lasting convention in SMT literature, where the letters f and e became the standard variables to denote source ($f$rench) and target ($e$nglish) language strings in equations, regardless of the actual languages involved. Our work builds on the modern legacy of these corpus-based approaches. Specifically, we contribute to the OLDI Seed Corpus (NLLB Team et al., 2022; Maillard et al., 2023), a resource designed to bootstrap translation capabilities for low-resource languages. The OLDI shared task has spurred several recent efforts to expand this dataset, including the creation of Spanish (Cols, 2024), Italian (Ferrante, 2024; Haberland et al., 2024), and Bangla (Ahmed et al., 2024) partitions, each contributing to this growing ecosystem of open, multiparallel data.

6 Conclusion

In this paper, we have presented our contribution to the WMT 2025 OLDI shared task: a high-quality, human-post-edited French partition of the OLDI Seed Corpus. We have detailed our data creation methodology, which relied on a diverse array of MT systems and approaches—from traditional NMT models to various LLM prompting strategies—and a custom-built interface to facilitate an efficient and robust post-editing workflow. Our experimental validation, using the state-of-the-art MetricX-24 QE metric, confirmed that our final, manually-refined corpus is of significantly higher quality than any of the individual machine-generated hypotheses. Our primary motivation for this work is to provide a reliable pivot resource to enable the future development of translation capabilities for the under-resourced regional languages of France. We hope that this dataset, along with the supplementary collection of raw MT outputs, will serve as a valuable contribution to the community and a stepping stone towards this long-term goal.

Limitations

Our discussion of relative system performance in Section 4 should be interpreted with one caveat. To enable qualitative observations during post-edition, we did not anonymize the MT systems in our interface (see Figure 1), a design choice that may have led to a "halo effect." The particular strength of one system (DeepSeek-R1) on challenging terminology, for instance, might have favorably influenced post-editor choice over time and slightly biased our logs.[8]

A key challenge in this work was navigating the tension between the OLDI translation guidelines, which tend to favor a more literal translation approach, and the need to resolve the stylistic and grammatical disfluencies often present in the source Wikipedia segments. While our post-editing process aimed to produce fluent French, a review of the final data in isolation from the source text reveals that some of these disfluencies, while corrected, may have still carried over into the target segments to some extent. This appears to be a common difficulty when working with this particular source corpus; indeed, a brief review of the recently released Italian (Ferrante, 2024; Haberland et al., 2024) and Spanish (Cols, 2024) partitions suggests that their respective translation teams grappled with similar issues. As this French corpus is intended primarily as a pivot resource, we plan to monitor its use in the translation into regional languages of France and will consider releasing future revisions should any significant issues be surfaced during this process.

Ethics Statement

In adherence to the OLDI shared task's commitment to open data, all systems used to generate translation hypotheses were carefully selected to ensure that their terms of service were compatible with the final dataset's release under a CC BY-SA 4.0 license. While we initially considered including hypotheses from popular commercial systems such as Google Translate and DeepL, we ultimately decided against it, as their terms of service prohibit the use of their outputs for the purpose of training other machine translation models.

The generation of translation hypotheses using LLMs is a computationally intensive process with an associated environmental cost. We strove to be mindful of this fact throughout our work.

We hope that our contribution will encourage further community-driven expansions of the OLDI Seed Corpus. As human societies grapple with the threat of "digital language death" (Kornai, 2013), the OLDI Seed Corpus project is particularly valuable. Its focus on the idiosyncratic Wikipedia domain is strategic, as the online encyclopedia plays a dual role in language revitalization: it is both a direct resource for creating the NLP artifacts essential for technological support, and a key instrument for "language ascent" (Kornai, 2013) that helps communities bridge the digital divide. Expanding this multiparallel dataset is therefore a direct and meaningful way to support these vital efforts.

Acknowledgements

This work was funded by the French *Agence Nationale de la Recherche* (ANR) under the project TraLaLaM ("ANR-23-IAS1-0006"), as well as by Inria under the "*Défi*"-type project COLaF. The last two authors' participation was also partly funded through their chairs in the PRAIRIE institute, supported by the ANR as part of the "*Investissements d'avenir*" program under the reference "ANR-19-P3IA-0001," and through the "*France 2030*" program under the reference "ANR-23-IACL-0008."

This work was also granted access to the HPC resources of IDRIS under the allocation 2025-AD011015117R2 made by GENCI.

[8] Reassuringly, however, the initial listing of translation candidates in decreasing order of their COMET QE scores in the post-editing interface does not appear to have induced a strong bias of its own. The average ranks for the top seven systems were tightly clustered in a narrow range (4.12–4.44 out of 9), indicating that no one system consistently dominated the top of the list, suggesting that post-editors were presented with a varied set of top-ranked candidates for most segments. While DeepSeek-R1 was the system whose output was most frequently selected by post-editors as "perfect" (requiring no edits in 40.42% of cases), the automated metric ranked it first for only 23.53% of segments, and its average rank of 4.27 was in the middle of the pack. This discrepancy supports the hypothesis that the post-editors' preference was driven by a human assessment of quality (particularly on terminology) rather than a bias induced by the tool's ranking.

References

Firoz Ahmed, Nitin Venkateswaran, and Sarah Moeller. 2024. The Bangla/Bengali Seed Dataset Submission to the WMT24 Open Language Data Initiative Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 556–566, Miami, Florida, USA. Association for Computational Linguistics.

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett,

John D. Lafferty, Robert L. Mercer, Harry Printz, and Lubos Ures. 1994. The Candide System for Machine Translation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Nicola Bertoldi and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece. Association for Computational Linguistics.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Ian Collen and Jayne Duff. 2025. Language Trends England 2025: Language teaching in primary, secondary and independent schools in England. Survey report, British Council.

Jose Cols. 2024. Spanish Corpus and Provenance with Computer-Aided Translation for the WMT24 OLDI Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 624–635, Miami, Florida, USA. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

H. P. Edmundson and D. G. Hays. 1958. Research methodology for machine translation. *Mechanical Translation*, 5(1):8–15.

Edoardo Ferrante. 2024. A High-quality Seed Dataset for Italian Machine Translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 567–569, Miami, Florida, USA. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. Italian-Ligurian Machine Translation in its Cultural Context. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages @ LREC-COLING 2024*, pages 168–176.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

András Kornai. 2013. Digital Language Death. *PLOS ONE*, 8(10):1–11.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Natalia Lusin, Terri Peterson, Christine Sulewski, and Rizwana Zafer. 2023. Enrollments in Languages Other Than English in US Institutions of Higher Education, Fall 2021. Report, Modern Language Association of America, New York, NY.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

OIF. 2022. *La langue française dans le monde 2022*. Éditions Gallimard. Préface de Souleymane Bachir Diagne.

Joseph Olive. 2005. Global autonomous language exploitation (GALE). *DARPA/IPTO Proposer Information Pamphlet*.

OST. 2024. The scientific position of France in the world and in Europe: Analysis of various corpora of publications and European projects. Prepared under the direction of Frédérique Sachwald; with contributions by Mounir Amdaoud, Agénor Lahatte, Esther Lardreau, and Françoise Laville.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Peter Rickard. 2014. *A History of the French Language*. Routledge. Google-Books-ID: kHm3oAEACAAJ.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Prompt Samples

A.1 Segment-Level Prompt

> You are an expert English-French translator of encyclopedic documents. In translating, you adhere to the following guidelines:
> 1. Refer to the source document context when available. Context helps clarify meaning, resolve ambiguities, and maintain tone and accuracy in translation.
> 2. Do not convert any units of measurement. Translate them exactly as noted in the source content.
> 3. Encyclopedic documents should be translated using a formal tone.

4. Provide fluent translations without deviating excessively from the structure of the source segment.
5. Do not expand or replace information compared to what is present in the source segment. Do not add any explanatory or parenthetical information, definitions, etc.
6. Do not ignore any meaningful text that was present in the source segment.
7. If a named entity in the source language has a canonical equivalent in the target language, use this canonical equivalent.
8. If a named entity in the source language does not have a canonical equivalent in the target language, you may use the source term in your translation.
You are to translate the following English source segment into French:
"Her father died in Norman, Oklahoma, in 1912, but she had returned to Ohio a few months before this."
Here is the segment in some of its original context; please note that the context may include parts of the source document that are not directly adjacent to the main segment, and omissions may not be explicitly marked with ellipses. Nonetheless, this context remains valuable for clarifying meaning, resolving ambiguity, and ensuring consistency in tone and terminology:
Gish was a prominent film star from 1912 into the 1920s, being particularly associated with the films of director D. W. Griffith.
She also did considerable television work from the early 1950s into the 1980s, and closed her career playing opposite Bette Davis in the 1987 film The Whales of August.
The first several generations of Gishes were Dunkard ministers.
Their mother opened the Majestic Candy Kitchen, and the girls helped sell popcorn and candy to patrons of the old Majestic Theater, located next door.
The seventeen-year-old Lillian traveled to Shawnee, Oklahoma, where James's brother Alfred Grant Gish and his wife, Maude, lived.
Her father died in Norman, Oklahoma, in 1912, but she had returned to Ohio a few months before this.
A reminder that the English segment you must translate into French is:
"Her father died in Norman, Oklahoma, in 1912, but she had returned to Ohio a few months before this."
You may reflect on the task at hand and explain your chain of thought prior to producing the translation.
IMPORTANT: Do write your translation between tags in the following manner: <translation>your translation here</translation>.

A.2 Document-Level Prompt

You are an expert English-French translator of encyclopedic documents. In translating, you adhere to the following guidelines:
1. Refer to the source document context when available. Context helps clarify meaning, resolve ambiguities, and maintain tone and accuracy in translation.
2. Do not convert any units of measurement. Translate them exactly as noted in the source content.
3. Encyclopedic documents should be translated using a formal tone.
4. Provide fluent translations without deviating excessively from the structure of the source segment.
5. Do not expand or replace information compared to what is present in the source segment. Do not add any explanatory or parenthetical information, definitions, etc.
6. Do not ignore any meaningful text that was present in the source segment.
7. If a named entity in the source language has a canonical equivalent in the target language, use this canonical equivalent.
8. If a named entity in the source language does not have a canonical equivalent in the target language, you may use the source term in your translation.
You are to translate the following English document into French. Please note that the following document may include omissions that are not explicitly marked with ellipses. Do not be perturbed by such minor inconsistencies in the source text. These segments were taken from an English Wikipedia page dedicated to 1. IMPORTANT! Please take careful note of the newline characters, as you will need to reproduce them perfectly in your French translation to allow for the automatic alignment of these segments with their English source.
—- BEGINNING OF ENGLISH SOURCE DOCUMENT —-
1 (one, also called unit, and unity) is a number and a numerical digit used to represent that number in numerals.
In conventions of sign where zero is considered neither positive nor negative, 1 is the first and smallest positive integer.
Most if not all properties of 1 can be deduced from this.
It is thus the integer after zero.
It was transmitted to Europe via the Maghreb and Andalusia during the Middle Ages, through scholarly works written in Arabic.
Styles that do not use the long upstroke on digit 1 usually do not use the horizontal stroke through the vertical of the digit 7 either.
By definition, 1 is the magnitude, absolute value, or norm of a unit complex number, unit vector, and a unit matrix (more usually called an identity matrix).
In category theory, 1 is sometimes used to denote the terminal object of a category.
Since the base 1 exponential function (1x) always equals 1, its inverse does not exist (which would be called the logarithm base 1 if it did exist).
Likewise, vectors are often normalized into unit vectors (i.e., vectors of magnitude one), because these often have more desirable properties.
It is also the first and second number in the Fibonacci sequence (0 being the zeroth) and is the first number in many other mathematical sequences.
Nevertheless, abstract algebra can consider the field with one element, which is not a singleton and is not a set at all.
A binary code is a sequence of 1 and 0 that is used in computers for representing any kind of data.
+1 is the electric charge of positrons and protons.
The Neopythagorean philosopher Nicomachus of Gerasa affirmed that one is not a number, but the source of number.
We Are Number One is a 2014 song from the children's TV show LazyTown, which gained popularity as a meme.
In association football (soccer) the number 1 is often given to the goalkeeper.

1 is the lowest number permitted for use by players of the National Hockey League (NHL); the league prohibited the use of 00 and 0 in the late 1990s (the highest number permitted being 98).
—- END OF ENGLISH SOURCE DOCUMENT —-
Only output the translation directly, religiously respecting new lines. Do not add extraneous new lines.

A.3 Document-Level Prompt with Corresponding French Wikipedia Article

You are an expert English-French translator of encyclopedic documents. In translating, you adhere to the following guidelines:
1. Refer to the source document context when available. Context helps clarify meaning, resolve ambiguities, and maintain tone and accuracy in translation.
2. Do not convert any units of measurement. Translate them exactly as noted in the source content.
3. Encyclopedic documents should be translated using a formal tone.
4. Provide fluent translations without deviating excessively from the structure of the source segment.
5. Do not expand or replace information compared to what is present in the source segment. Do not add any explanatory or parenthetical information, definitions, etc.
6. Do not ignore any meaningful text that was present in the source segment.
7. If a named entity in the source language has a canonical equivalent in the target language, use this canonical equivalent.
8. If a named entity in the source language does not have a canonical equivalent in the target language, you may use the source term in your translation.
The document you will translate consists in segments taken from an English Wikipedia page dedicated to North. Here is what appears to be the corresponding French Wikipedia page (back-matter sections have been removed). It might provide you with the correct terminology and equivalent named entities, pay close attention to these aspects as you read this French text.
—- BEGINNING OF FRENCH WIKIPEDIA ARTICLE —-
Le nord est un point cardinal, opposé au sud.
== Étymologie ==
De l'ancien haut-allemand nord provenant de l'unité linguistique proto-indo-européenne « ner- » qui signifie « gauche », se rapportant sans doute à la gauche du soleil levant.
Le nom de la divinité scandinave Njörd, ayant régné sur une partie du monde pendant un âge d'or, est lié à cette racine[réf. souhaitée]. Cette divinité était connue des Romains sous le nom de Nerthus et avait donné son nom à une des îles du bout du monde, Nérigon.
En latin, Septemtriones signifie les sept bœufs. L'astérisme le plus brillant de l'actuelle constellation de la Grande Ourse, était autrefois une constellation à part entière appelée constellation des sept bœufs. Ce groupement d'étoiles permettait de trouver l'étoile polaire et donc le Nord avec une bonne précision.
Le terme septentrion est un synonyme vieilli de nord, faisant référence à cette constellation qui indiquait la direction du nord aux Romains ; mais l'adjectif septentrional, qui en découle, reste très usité.
== Géographique et magnétique ==
Il existe deux nord. Le premier est magnétique (l'axe de symétrie cylindrique du champ magnétique), le second est géographique (l'axe de rotation de la Terre). Ces deux points ne se trouvent pas au même endroit. Mesuré en 2007 par le projet « Poly-Arctique », le pôle Nord magnétique est situé à 83° 57' 00" N, 121° 01' 12" O. Il se trouve à 673 km du pôle Nord géographique et ayant une vitesse moyenne de déplacement de 55 km/an (soit une moyenne d'environ 150 m/jour ou 6 m/h). À l'été 2010, il a été estimé qu'il n'était plus qu'à 550 km du pôle Nord géographique.
La différence d'angle que l'on peut observer sur la boussole entre ces deux nord est appelée déclinaison magnétique. Cette différence varie avec le temps.
Sur les cartes traditionnelles et en particulier les cartes de l'Institut national de l'information géographique et forestière (IGN), les méridiens (lignes noires verticales) pointent le nord géographique (NG) ; il y a donc lieu de tenir compte de la déclinaison magnétique pour s'orienter sur la carte à l'aide d'une boussole (NM). Le croquis situé dans la légende de la carte indique la valeur de la déclinaison pour la carte et pour une année donnée, car le pôle magnétique migre en permanence, réduisant chaque année la valeur de la déclinaison (0,8 degré/an).
Certains cartographes ont contourné cette complication en construisant des cartes tenant compte de cette déclinaison : le nord (N) de la carte ainsi que les lignes verticales en bleu ou en noir pointent le nord magnétique (de la même manière que l'aiguille de la boussole).
La position du nord magnétique a changé plusieurs fois dans l'histoire de la Terre ; la dernière inversion du champ magnétique terrestre s'est produite il y a 780 000 ans.
En l'absence de boussole, le moyen traditionnel pour repérer le nord le soir ou la nuit est de se référer à l'étoile polaire dans l'hémisphère nord ou à la croix du Sud dans l'hémisphère sud. Le jour, il est possible de se référer à la position du Soleil en fonction de l'heure locale. Lorsque le ciel est couvert, observer la mousse ou les vents dominants est peu fiable.
== Typographie ==
Les points cardinaux, qu'ils soient utilisés comme nom ou comme qualificatif, s'écrivent avec :
une majuscule lorsqu'ils font partie d'un toponyme ou désignent une région ;
une minuscule s'ils désignent une direction, une exposition, une orientation.
=== Articles connexes ===
Sud
Point cardinal
—- END OF FRENCH WIKIPEDIA ARTICLE —-
You are to translate the following English document into French. Please note that the following document may include omissions that are not explicitly marked with ellipses. Do not be perturbed by such inconsistencies in the source text. IMPORTANT! Please take careful note of the newline characters, as you will need to reproduce them perfectly in your French translation to allow for the automatic alignment of these segments with their English source.
—- BEGINNING OF ENGLISH SOURCE DOCUMENT (this line need not be translated) —-
North is one of the four compass points or cardinal directions.

Septentrionalis is from septentriones, "the seven plow oxen", a name of Ursa Major.
For example, in Lezgian, kefer can mean both "disbelief" and "north", since to the north of the Muslim Lezgian homeland there are areas formerly inhabited by non-Muslim Caucasian and Turkic peoples.
On any rotating astronomical object, north often denotes the side appearing to rotate counter-clockwise when viewed from afar along the axis of rotation.
But simple generalizations on the subject should be treated as unsound, and as likely to reflect popular misconceptions about terrestrial magnetism.
—- END OF ENGLISH SOURCE DOCUMENT (this line need not be translated) —-
IMPORTANT! Only output the translation directly, religiously respecting new lines. Do not add extraneous new lines. Do not skip any segment.

Bringing Ladin to FLORES+

Samuel Frontull[1], Thomas Ströhle[1], Carlo Zoli[2], Werner Pescosta[3], Ulrike Frenademez[3], Matteo Ruggeri[3], Daria Valentin[3], Karin Comploj[3], Gabriel Perathoner[3], Silvia Liotto[3], Paolo Anvidalfarei[3]

[1]Department of Computer Science, University of Innsbruck, Austria
[2]Faculty of Education, Free University of Bozen-Bolzano, Italy
[3]Ladin Cultural Institute "Micurá de Rü", San Martin de Tor, Italy

Correspondence: samuel.frontull@uibk.ac.at

Abstract

Recent advances in neural machine translation (NMT) have opened new possibilities for developing translation systems also for smaller, so-called low-resource, languages. The rise of large language models (LLMs) has further revolutionized machine translation by enabling more flexible and context-aware generation. However, many challenges remain for low-resource languages, and the availability of high-quality, validated test data is essential to support meaningful development, evaluation, and comparison of translation systems. In this work, we present an extension of the FLORES+ dataset for two Ladin variants, Val Badia and Gherdëina, as a submission to the Open Language Data Initiative Shared Task 2025. To complement existing resources, we additionally release two parallel datasets for Gherdëina–Val Badia and Gherdëina–Italian. We validate these datasets by evaluating state-of-the-art LLMs and NMT systems on this test data, both with and without leveraging the newly released parallel data for fine-tuning and prompting. The results highlight the considerable potential for improving translation quality in Ladin, while also underscoring the need for further research and resource development, for which this contribution provides a basis.

1 Introduction

In recent years, the field of machine translation (MT) and natural language processing has advanced rapidly and the transformer-based models played a key role in this development (Vaswani et al., 2023; Aharoni et al., 2019). This paradigm shift has enabled the development of high-quality MT systems for major languages as well as the adaptation of such systems to low-resource languages by leveraging pre-trained knowledge (Zoph et al., 2016; Kocmi and Bojar, 2018). Earlier rule-based and statistical MT approaches lacked this capability as they depended on large, clean, and domain-specific parallel corpora, thus limiting the availability of these technologies for most low-resource languages. The release of large-scale multilingual systems, such as No Language Left Behind (NLLB) (NLLB Team et al., 2024), has demonstrated that scalable and accurate translation is possible even for smaller languages.

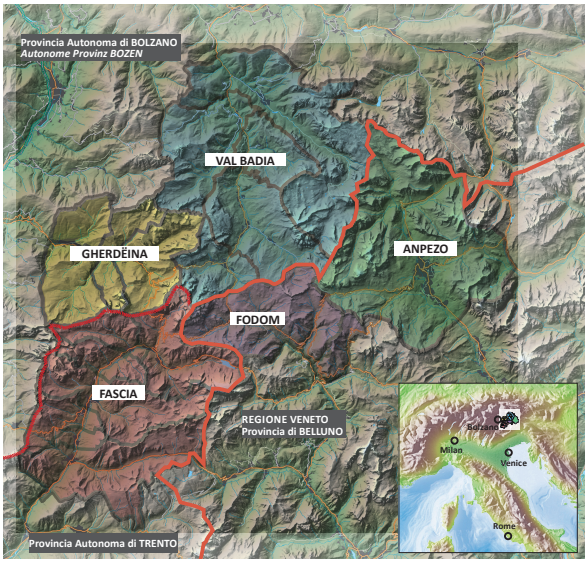


Figure 1: The five Ladin-speaking valleys located across three provinces in northern Italy: Val Badia (blue) and Gherdëina (yellow) in the Autonomous Province of Bolzano/Bozen South Tyrol, Fascia (red) in the Autonomous Province of Trento, Fodom (purple) and Anpezo (green) in the Province of Belluno.

Essentially, developing machine translation systems has actually become easier as it is no longer just about language expertise. Instead, the generalizability capabilities of language models are being exploited, and the limitations now lie more in providing training data and infrastructure. However, its development depends on reliable and extensive test data, which is essential for consistent evaluation, model comparison, and tracking progress over time. The FLORES+ dataset provides such a benchmark and covers more than 200 languages (NLLB Team et al., 2024). It is widely used in research on

Proceedings of the Tenth Conference on Machine Translation, pages 1061–1071
November 8-9, 2025

low-resource languages because of its broad language coverage and its ability to support evaluation across more than 19,900 language pairs.

With this in mind, we release the FLORES+ translations for the Ladin language. Ladin is a Rhaeto-Romance language that originated from Vulgar Latin during the Roman conquest of the Alps. It evolved in isolated Alpine valleys, leading to the diverse regional varieties spoken today in Northern Italy (Bauer, 2022). As a result of the fragmented development over centuries there is no single, unified spoken variety for Ladin that is characterised by its internal linguistic diversity, with five main regional variants depicted in Figure 1: *Val Badia*, *Gherdëina*, *Fassa*, *Fodom*, and *Anpezo*. In this work, we extend FLORES+ with translations for Val Badia and Gherdëina. Both variants are spoken in South Tyrol and together representing the largest portion of Ladin speakers. The dataset translation was carried out by professionals at the Ladin Cultural Institute "Micurá de Rü", the primary institution dedicated to preserving and supporting the development of the language.

Moreover, we release parallel datasets for Val Badia–Gherdëina and Gherdëina–Italian that can be used for fine-tuning or retrieval-augmented prompting for this language. These two datasets enables to provide an overview of the performance of `NLLB` and state-of-the-art Large Language Models (LLMs), including `GPT-3.5`, `GPT-4o`, `Llama-3.3`, and `DeepSeek-R1` on Ladin (Val Badia and Gherdëina).

In summary, the contributions of our work are:

1. We submit the FLORES+ translations for Val Badia Ladin (full) and Gherdëina Ladin (`dev` split)

2. We release two parallel datasets of approximately 18,000 sentence pairs for Gherdëina–Italian and Gherdëina–Val Badia.

3. We benchmark state-of-the-art machine translation systems and LLMs on this dataset, identifying current limitations and outlining opportunities for advancing MT for Ladin.

With this, we aim to increase the visibility of Ladin within the MT research community, highlight the ongoing need for focused research on this language, and provide resources that empower researchers and developers worldwide to advance Ladin machine translation.

2 The Ladin Language

Ladin should not be confused with *Ladino* (`lad`), a Judeo-Spanish language. This confusion is common, especially in Italian where Ladino can refer to both; to avoid ambiguity, the term *Ladino delle Dolomiti* (Dolomite Ladin) is often used. Ladin is identified by the ISO 639-3 code `lld`.

For a comprehensive treatment of the language's history, dialectal variation, and sociolinguistic context, we refer the reader to Pescosta (2015); Videsott et al. (2020); Bauer (2022). For a recent contribution discussing the state of the Ladin language in the Dolomite region, we refer to Videsott (2023); Colcuc (2024).

2.1 Historical Development

Ladin is a Romance language that is traditionally associated with Romansh of the Grisons in Switzerland and Friulian of north-eastern Italy within the Rhaeto-Romance group (Videsott et al., 2020). It traces its origins to the Vulgar Latin introduced during the Roman conquest of the Alpine region around 15 BC. Over time, the native populations gradually adopted this Latin, which absorbed elements of pre-Roman languages such as Celtic and Raetic. With the fall of the Western Roman Empire, the evolving Romance dialects began to diverge, influenced by the surrounding Germanic languages and political fragmentation. As Germanic tribes advanced, the speakers of early Ladin were pushed into isolated Alpine valleys, where the language continued to evolve separately from other Romance varieties. This geographic and historical isolation laid the foundation for the internal diversity of Ladin seen today in the Dolomite region of Northern Italy where each variant reflects the unique historical trajectory and degree of contact with neighbouring languages.

The internal linguistic diversity of Ladin has not only historical roots but has also been reinforced by political-administrative fragmentation in the 20th century. Following the annexation of South Tyrol by Italy after the Treaty of Saint-Germain (1919) and the Italianisation policies introduced under the Fascist regime, Ladin-speaking territories were divided among three different provinces: *Val Badia* and *Gherdëina* became part of the autonomous province of Bolzano (South Tyrol), *Fassa* was integrated into the province of Trento, and *Fodom* and *Anpezo* were assigned to the province of Belluno

in the Veneto region (see Figure 1[1]). This fragmentation led to differing levels of legal recognition, educational support, and public visibility for Ladin communities depending on the province.

2.2 Contemporary Situation

The degree of institutional support and practical implementation of Ladin varies among the different regions. In South Tyrol, Ladin enjoys strong official recognition and robust backing in administration, education, and media. According to the 2024 South Tyrol census[2], approximately 4.41% (19.853) of South Tyroleans declared themselves as belonging to the Ladin language group. The highest proportions of Ladin speakers are found in Val Badia, where 97% of the population, and in Val Gardena, where 85% of the population, identify as Ladin-speaking. This corresponds to around 20,000 people Ladin speakers in South Tyrol.

In the Trentino (Fassa Valley), Ladin is also officially recognized but coexists with a stronger presence of Italian, resulting in comparatively less institutional support and daily use. Based on the 2021 census[3], 2.9% (15.775) of residents identified as Ladin speakers in Trentino.

Meanwhile, in the Veneto valleys of Livinallongo and Ampezzo, its status is more limited, and the language faces greater challenges regarding vitality and institutional presence. The precise number of native Ladin speakers in the traditional (formerly Austrian) Dolomites Ladin area of Veneto (namely Fodom and Anpezo valleys) is not available, but can be estimated to be around 4-6,000[4]. Together, these figures amount to an estimated 40.000 Ladin speakers.

Three cultural institutes, set up as public bodies in the three different Provinces, with a total staff of around 20 people[5] work on standardisation, the creation of digital tools, and the promotion of the social status of the language, following the three traditional axes of language policy for minority languages (status-, corpus- and acquisition planning; see Iannaccaro and Dell'Aquila (2004) for further details). Moreover, a Chair of Ladin Language and Culture Studies has been established at the Free University of Bolzano (*Université Lieida de Bulsan*)[6].

2.3 Linguistic and MT Research on Ladin

Linguistic research on Ladin has a long tradition and remains very active, notably through the journal *Ladinia*[7], which has appeared 49 times since its inception in 1977. Another journal devoted to Ladin studies is *Mondo Ladino*[8], also founded in 1977.

While few published parallel datasets exist for Ladin, existing experiments indicate that state-of-the-art machine translation techniques can be adapted to this low-resource language. Frontull and Moser (2024b) developed an NMT system for Val Badia Ladin and compared it to rule-based and statistical systems; Frontull and Moser (2024a) studied back-translation using `GPT-3.5`; and Valer et al. (2024) created a bidirectional system for Fassa Ladin using multilingual training and knowledge transfer, comparing it to `GPT-4o`. However, the test data used in these studies was limited in scope, preventing broader comparability of results and restricting evaluation to Ladin–Italian translation. There is still much potential to explore how to make even better use of the available resources.

2.4 Ladin of Val Badia and Gherdëina

Although ISO 639-3 assigns a single code (`lld`) to Ladin, it may be more accurate to consider `lld` as a macrolanguage encompassing at least five standardized written variants: Ladin of *Val Badia*, *Gherdëina*, *Fascian*, *Fodom*, and *Anpezan*. In this work, we focus on and provide translations for the two written standards of Val Badia Ladin and Gherdëina Ladin, which can be specifically identified using the IETF BCP 47 language tags[9] `lld_valbadia` and `lld_gherd` respectively. These correspond to Glottolog code `gard1241`[10] for Gherdëina. For Val Badia, there is no exact correspondence, as the spelling unifies `badi1244`

[1] An interactive map is available at https://atlantilinguistici.smallcodes.com/ladinia.html

[2] https://astat.provinz.bz.it/de/publikationen/ergebnisse-sprachgruppenzahlung-2024

[3] http://www.statistica.provincia.tn.it/binary/pat_statistica_new/popolazione/Sintesi_Rilevazione_minoranze_2021.1651135663.pdf

[4] a conservative estimate informed by anecdotal evidence, calculated for a total population of 7,000

[5] a comparatively large number given the size of the population

[6] https://www.unibz.it/

[7] https://www.micura.it/de/ativites-2/ladinia, ISSN 1124-1004

[8] https://www.istladin.net/en/pubblicazioni, ISSN 1121-1121

[9] https://www.iana.org/assignments/language-subtag-registry

[10] https://glottolog.org/resource/languoid/id/gard1241

English	The walls and roofs of ice caves can collapse and cracks can get closed.
Italian	Le pareti e il soffitto delle caverne di ghiaccio sono soggetti a crolli e le crepe possono richiudersi.
Val Badia	I parëis y le sössot di andri da dlacia pó tomé ite y les sfësses pó se stlüje pro.
Gherdëina	I parëies y i plafons dla ciavernes tla dlacia possa tumé ite y la sfëntes possa se stlù.

Table 1: Example translations from the FLORES+ dev split.

and mare1258, with badi1244[11] being the more appropriate mapping.

Linguistic Features Socio-linguistically, in Gardena and Badia valleys Ladin is in contact both with German (in the local dialectal form and in the standard official form) and Italian, in the remaining valleys Italian is the dominant contact language (Videsott et al., 2020). Being Ladin a romance language, standard Italian is in any case a natural point of comparison to highlight important linguistic differences that affect machine translation. Ladin differs from Italian in several ways that are relevant for MT. In the following, we summarize key differences taken from (Bauer, 2022) Phonologically, Ladin features voicing of intervocalic stops (e.g., Latin *p*, *t*, *k* become voiced between vowels), retention of final *-s* to mark plurals, and palatalisation of *c*, *g* before *a*. These sound changes are consistently reflected in Ladin spelling, which means the written forms differ systematically from Italian. Morphologically, Ladin has a complex pattern of plural endings (*-s*, *-i*, other form of palatised consontant) that Italian lacks. Additionally, Ladin has a high number of loanwords from Germanic languages due to long-standing contact (especially for northern varieties), with southern varieties borrowing mainly from Italian, resulting in significant lexical variation even within Ladin written and oral varieties. These phonological, morphological, and lexical differences shape the vocabulary and structure that MT systems must handle, making their accurate representation essential for building effective Ladin machine translation models and a compelling case for natural language processing research.

To give an intution on the similarity between the two variants and Italian and on the difficulty of the translation task, we computed the BLEU score obtained by leaving the text untranslated, which resulted in a score of 5.0 for Italian–Val Badia, 4.3 for Italian–Gherdëina and 12.9 for Val Badia–Gherdëina. Table 1 presents a sample translation of the English sentence *"The walls and roofs of ice caves can collapse and cracks can get closed."* into Val Badia and Gherdëina, as found in the submitted dev split of FLORES+, illustrating these similarities.

Relevant Resources Although relatively little previous research has focused on machine translation for Ladin, significant and valuable work has been carried out in the development of the language itself. Comprehensive dictionaries have been compiled, alongside essential language reference materials, including books detailing grammar and spelling rules. In the following we list relevant resources for the variants in focus in this work developed by the Ladin Cultural Institute "Micurá de Rü":

- *Grafia nöia - Ladin scrit dla Val Badia* (Mischí et al., 2015) is a practical guide to the standardized orthography of Ladin as used in Val Badia which was reformed in 2015.
- *La ortografia dl ladin de Gherdëina* (Forni, 2019b) is a practical guide to the standardized orthography of Ladin as used in Val Gardena.
- *Dizionario italiano - ladino Val Badia* (Moling et al., 2016) is a bilingual dictionary published in 2016. The dictionary includes 30,829 Italian lemmas and 32,701 Ladin lemmas, along with 18,120 phraseological expressions. It offers morphological and encyclopedic information.
- *Dizionario italiano - ladino gardenese* (Forni, 2013) is a bilingual dictionary published in 2013. Authored by Marco Forni, this work is an essential resource for the Ladin language as spoken in Val Gardena (Gherdëina). This extensive bilingual dictionary, contains over 67,000 entries and nearly 20,000 phraseological expressions, provides detailed lexical information along with contextual examples.
- *Gramatica Ladin Gherdëina* (Forni, 2019a), authored by Marco Forni and published in

[11] https://glottolog.org/resource/languoid/id/badi1244

2019. Covering phonetics, morphology, and syntax, it is a comprehensive grammatical reference for the Ladin language as spoken in Val Gardena.

- Several *technical and domain-specific glossaries* are available for Ladin, including, for example, mobile phone interfaces fully translated in collaboration with Motorola (Oliveira et al., 2024) and terminology in pedagogy[12]. Glossaries in other domains, such as music, animals and plants, history, and historiography, are currently under elaboration.

- *Spellchecker*: online tools[13] as well as Firefox Add-ons are available for Val Badia[14] and Gherdëina. [15]

- set of *18k parallel sentences for Ladin Val Badia–Italian*, as used in (Frontull and Moser, 2024a), available at HuggingFace[16]

We leveraged the existing dictionaries to implement spelling correction and extracted the parallel sentences contained in the datasets we release these resources.

3 Data collection

The translation of the FLORES+ dataset into Ladin was carried out in close collaboration with the Ladin Cultural Institute "Micurá de Rü"[17]. In this section, we provide an overview of the translators involved, the tools used, the procedures followed for the different Ladin variants, and the quality assurance measures implemented.

Translators The translation team consisted of five translators for Val Badia and two translators for Gherdëina. All translators involved in this process are (or were at the time of the project) employed by the Ladin Cultural Institute, where they work professionally with Ladin and therefore have extensive experience using the language in both spoken and written forms. They are native speakers of Ladin and hold a C1-level certification in Ladin, obtained through the official language exams[18] regulated by the Autonomous Province of Bolzano/Bozen, making them highly valuable collaborators, whose contributions ensure reliable, high-quality work. All translators also have a good understanding of English (though not formally certified) and are fluent at C1 level in both German and Italian. Prior to beginning the task, all translators were asked to carefully read and acknowledge the OLDI translation guidelines[19], which explicitly require human translation, prohibit the use of machine translation systems and provide detailed instructions on tone, consistency, fidelity to the source, and the handling of named entities. In accordance with these guidelines, translators were instructed to base their work primarily on the English source texts. However, due to their proficiency in German and Italian, these translations were also provided as additional references. Due to the linguistic similarities between Ladin and Friulian, these texts were also included as reference texts.

Translation Assignment Tool The work was divided into kits of 25 sentences. To manage and assign the translation work, a dedicated web-based platform was set up. This tool allowed kits to be assigned to specific translators and the progress of the translation to be monitored, eliminating the need to distribute and manage separate text files. The tool displays the original English text along with reference translations into German, Italian, Friulian, and, for Gherdëina, also the translation into Val Badia. Translators can enter their Ladin translation directly into an input field. The tool also has a built-in spell checker that highlights unknown or potentially misspelled words and helps users identify and correct typos. Figure 2 shows a screenshot of the user interface for a sentence to be translated, with the English source text, reference translations, and the input field for the Ladin translation. Sentences can also be skipped and revisited later.

Val Badia Translations In a first phase, the sentences were translated into the Ladin Val Badia. Each translator received two kits per week, a work-

[12] https://pedagogia.ladinternet.it/glossary

[13] https://www.micura.it/en/online-services/spellchecker

[14] https://addons.thunderbird.net/de/thunderbird/addon/lld-valbadia/

[15] https://addons.thunderbird.net/de/thunderbird/addon/lld-gherd/

[16] https://huggingface.co/datasets/sfrontull/lld_valbadia-ita

[17] https://micura.it

[18] The exams include listening, writing, speaking, and reading comprehension and are offered in both the Gherdëina and Val Badia varieties. This certification is a prerequisite for a permanent employment at the Ladin Cultural Institute in South Tyrol. For more information we refer to https://zweisprachigkeitspruefungen.provinz.bz.it/de/ladinischpruefung

[19] https://oldi.org/guidelines

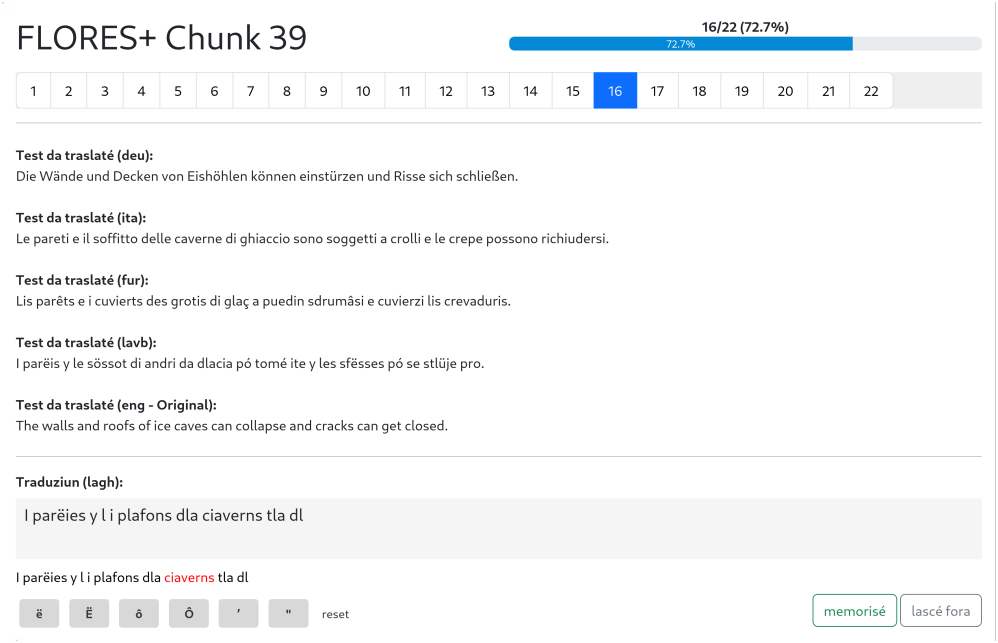


Figure 2: Screenshot of the tool developed for the translation of FLORES+.

load that proved manageable alongside daily tasks. This modular approach ensured an even distribution of work, allowed careful review of each unit, and maintained a sustainable pace. Overall, the team was able to complete the work within four months.

Gherdëina Translations Building on the Val Badia translations, we began translating into Ladin Gherdëina, using the Val Badia translations as an additional reference. This sequential approach was practical, as adapting an existing Ladin text is more efficient than translating from scratch, particularly with limited personnel. This second translation cycle also served as an implicit revision of the Val Badia texts, during which minor errors were identified and corrected. Given the smaller number of translators, the project timeline has been adjusted to ensure careful and thorough completion (work is still in progress).

Quality Assurance The first author, also a native Ladin speaker, continuously reviewed a sample of the generated translations and extracted all unknown words detected with a variant-specific spell checker. These were compiled into a shared document, where translators were asked to review and confirm or correct the entries. Words requiring concordance across variants were fed back to the group, ensuring consistency. This collaborative monitoring helped to identify and resolve potential errors. Concerns about the use of external machine translation systems (e.g., Google Translate or ChatGPT) do not apply in this setting. Major MT systems do not support Ladin, nor its specific variants, and ChatGPT performs poorly on Ladin (as evidenced in our evaluation results below). Any attempt to use such tools would have been immediately detectable, as they are incapable of generating variant-specific vocabulary and would trigger numerous alerts in the spell checker.

Neologisms The translation activity was accompanied by intensive collaboration among the translators. This collaborative process not only helped resolve stylistic and terminological questions but also led to the creation of new words that had not previously existed in the respective variety. As such, the project contributed to the lexical development of the Ladin language. In total, around 500 new words were created. Examples of such words are: *codificades* (encoded), *dejabilité* (disability), *surafolamënt* (overcrowding) or *triceratop*.

4 Experimental Validation

In this section, we provide an overview of the results achieved with a NMT system and different state-of-the-art LLMs on the manually created FLORES+ translations. The aim is to assess the current capabilities of these models in translating between Val Badia, Gherdëina and Italian, both in off-the-shelf zero-shot settings and using the accompanying parallel data as fine-tuning training

data for the NMT system and as retrieval data for the LLMs. Moreover, this serves as validation of the quality of the submitted datasets.

Training and Retrieval Corpora For Italian–Val Badia, a dataset is already available[20] consisting of word usage example sentences extracted from the Italiano–Ladin Val Badia dictionary (Moling et al., 2016). To create training and retrieval resources for Gherdëina, we extracted the word usage example sentences contained in the Italiano–Ladin Gherdëina dictionary (Forni, 2013). Since the two dictionaries are based on a common foundation and were largely coordinated with each other, they contain many common example sentences. This overlap allowed us to construct a Val Badia–Gherdëina parallel dataset by aligning identical Italian sentences. This yielded the following datasets for training (for `NLLB`) and retrieval (for LLMs):

- 18,140 sentences for Val Badia–Italian,
- 19,971 sentences for Gherdëina–Italian, and
- 14,953 sentences for Val Badia–Gherdëina.

Note that, given their purpose of illustrating word usage, the sentences in these corpora are relatively short and simple, with an average length of approximately 25 characters. We make both the Gherdëina–Italian[21] and Val Badia–Gherdëina[22] datasets publicly available under a `CC BY-NC-SA 4.0`[23] license.

4.1 Neural Machine Translation Models

For this evaluation, we tested the performance of the multilingual NMT system `facebook/nllb-200-distilled-600M` (NLLB Team et al., 2024). We evaluated this system off-the-shelf (without any fine-tuning) from Ladin to Italian and after being fine-tuned with the parallel datasets available for Val Badia–Italian, Gherdëina–Italian and Val Badia–Gherdëina mentioned above. We trained a single model covering all translation directions simultaneously using the `transformers` framework for 10 epochs with a batch size of 16 (which corresponded to approximately 101,930 optimization steps) and a maximum sequence length of 196 tokens. The validation loss steadily improved throughout training. Training was completed in 7 hours and 42 minutes on an NVIDIA RTX 4090 GPU.

4.2 Large Language Models

LLMs are proving to be increasingly valuable translators. A key advantage of LLMs is their flexibility in adapting to different specifications and contexts, which enables more targeted, application-specific use. Several techniques have also been studied for machine translation in low-resource languages, showing that LLMs can perform well in these settings and can often be further improved by providing additional data (Agrawal et al., 2023; Tang et al., 2024). Motivated by this, we evaluated different LLMs on Ladin to gain an overview of their performance and the results they can achieve. We evaluate the following four LLMs: (*i*) `GPT-3.5` is a general-purpose language model from OpenAI's GPT-3 series with 175B parameters, released in 2022 (Brown et al., 2020). (*ii*) `GPT-4o` a model by OpenAI, introduced in 2023, with 200B parameters and enhanced reasoning capabilities designed for complex problem-solving (Hurst et al., 2024). (*iii*) `Llama-3.3` is a text-only model by Meta AI, released in December 2024, featuring 70B parameters (Touvron et al., 2023). (*iv*) `DeepSeek-R1` is a reasoning-focused model by DeepSeek AI, introduced in January 2025, with 658B parameters (DeepSeek-AI et al., 2025). The models were prompted using the API services: OpenAI API[24] for `GPT-3.5` and `GPT-4o`, DeepSeek API[25] for `DeepSeek-R1`, and Together Inference API[26] for `Llama-3.3`. The hyperparameters were configured according to the default settings provided by each service.

Due to the very limited amount of machine-readable data available for Ladin it is likely that LLMs have minimal exposure to Ladin and are unaware of its internal variation.

4.2.1 Prompting Techniques

We applied two prompting methodologies in our experiments: *zero-shot* and *BM25-Retrieval*. The *zero-shot* method (Robinson et al., 2023; Gao et al., 2024; Hendy et al., 2023; Bawden and Yvon, 2023) relies solely on the model's pre-existing knowledge. The prompt directly instructs the model to translate sentence into the target, without providing explicit

[20] https://doi.org/10.57967/hf/1878
[21] https://huggingface.co/datasets/sfrontull/lld_gherd-ita
[22] https://huggingface.co/datasets/sfrontull/lld_valbadia-lld_gherd
[23] https://creativecommons.org/licenses/by-nc-sa/4.0
[24] https://platform.openai.com
[25] https://www.deepseek.ai/api
[26] https://www.together.ai

Model	low- to high-resource VB→IT	low- to high-resource GH→IT	high- to low-resource IT→VB	high- to low-resource IT→GH	low- to low-resource VB→GH	low- to low-resource GH→VB
zero-shot / base model						
GPT-3.5	19.35±2.09	18.52±2.04	4.91±1.23	6.73±1.18	10.52±1.41	10.73±1.53
GPT-4o	**22.90**±2.11	**23.03**±2.09	5.70±1.40	5.53±1.20	11.15±1.61	11.17±1.37
Llama-3.3	20.31±2.10	21.02±2.12	6.46±1.29	9.50±1.35	15.01±1.69	12.46±1.53
DeepSeek-R1	22.47±2.04	23.02±1.96	6.31±1.24	8.77±1.30	13.11±1.52	10.19±1.34
NLLB BM	14.86	12.49	-	-	-	-
BM25-Retrieval / fine-tuned						
GPT-3.5	26.68±2.03	20.95±1.69	9.64±1.27	8.22±1.19	15.95±1.49	16.00±1.68
GPT-4o	**29.47**±2.25	**23.51**±2.16	14.49±1.55	11.62±1.41	22.72±1.97	19.58±2.00
Llama-3.3	27.20±2.04	22.33±1.98	15.50±1.77	13.20±1.51	24.35±1.89	21.22±1.98
DeepSeek-R1	28.91±1.93	22.30±1.83	16.06±1.66	14.07±1.46	24.81±1.78	22.49±2.08
NLLB FT	21.73	17.80	**18.06**	**14.48**	**29.63**	**31.15**

Table 2: BLEU mean scores and confidence intervals for each model across six translation directions, split by method (*zero-shot* and *BM25-Retrieval / fine-tuned*).

Model	IT→VB	IT→GH	VB→GH	GH→VB
reference	78%	80%	80%	78%
GPT-3.5	53%	52%	55%	58%
GPT-4o	58%	58%	64%	65%
Llama-3.3	60%	62%	68%	66%
DeepSeek-R1	63%	64%	67%	68%
NLLB FT	**76%**	**73%**	**76%**	**79%**

Table 3: Average proportion of words in manually created ladin translations and those generated with BM25-Retrieval/fine-tuned NLLB model passing spellcheck.

translation examples or lexical guidance. This baseline approach tests the model's intrinsic understanding of Ladin syntax and vocabulary. To enhance translation quality beyond zero-shot capabilities, we employ *BM25-Retrieval* to provide the model with relevant in-context examples based on lexical similarity. BM25 (Robertson et al., 1995) is a probabilistic ranking method that estimates the relevance of documents to a given search query. This sparse retrieval method ranks examples based on lexical-level overlap with the input sentence, aiming to offer relevant in-context examples grounded in lexical similarity. It has been shown to be highly effective also for retrieving examples for MT with LLMs (Agrawal et al., 2023; Tang et al., 2024). We implemented this method using the `bm25s` Python package[27] (Lù, 2024) to select 30 translation examples from the corresponding retrieval corpus that we included in the prompt.

5 Results

Table 2 presents the BLEU scores for the six translation directions between Val Badia (VB), Gherdëina (GH), and Italian (IT). The upper half of the table reports results obtained with the models "off-the-shelf," that is, using zero-shot prompting for `GPT-3.5`, `GPT-4o`, `Llama-3.3`, and `DeepSeek-R1`, as well as the NLLB base model (BM). The lower half shows the results obtained when exploiting additional parallel data as retrieval data and to fine-tune the NLLB model (FT). Since the LLMs were evaluated on a subset of 175 sentences, we report the average BLEU (Post, 2018) scores together with the standard deviation[28] to provide a confidence interval for the results. The best score in each translation direction, when comparing the LLMs with NLLB, is highlighted in bold. In Table 3, we report the average portion of words in the generated translations that pass spellcheck. This should give an idea of the general lexical quality of the translations.[29] Also in this table we highlighted the best scores (LLMs vs. fine-tuned NLLB model) in bold.

[27] https://github.com/xhluca/bm25s

[28] Computed with `sacrebleu` (Post, 2018)

[29] Note that the reference translations do not achieve 100% spellcheck accuracy primarily because many named entities are absent from the dictionary and thus flagged as invalid. Therefore, this metric serves mainly as an expectation indicator rather than an absolute measure of correctness.

Interestingly, concerning the fine-tuned NLLB model, the relatively simple training datasets proved to be effective, especially considering that the benchmark data is substantially more complex. Despite the brevity and simplicity of the training samples, the model generated complete translations of the test samples. One peculiarity we observed is that the generated translations usually start with lowercase letters, reflecting the format of the training examples. Another limitation is that the model struggles with morphological variations: since most training examples contained words in their infinitive form (e.g., "I go home" and never "he went home"), the model is unable to generate correct inflections or conjugations. LLMs are less influenced by the simple translation examples and do not "compromise" their fluency when translating into Italian. For example, they also use the correct capitalization at the beginning of the sentence. When translating from Italian into Ladin, however, LLMs often lack the appropriate vocabulary, whereas the fine-tuned NLLB model adapts the vocabulary more effectively (see Table 3). In contrast, translation between Ladin variants is an easier task, as it mainly involves adapting the vocabulary. Here, the NLLB model already delivers satisfactory results and the difference between NLLB and LLMs is more pronounced, highlighting the challenge LLMs face in accurately adapting vocabulary.

From these results, three main observations can be drawn: (i) current LLMs exhibit limited coverage of Ladin variants, with translation into Ladin remaining a clear challenge; (ii) incorporating the parallel data released in this work yields substantial improvements in translation quality across models, but limitations remain in fluency and morphological variation due to the simplicity of the training examples; (iii) the relative advantage of the systems depends on the translation direction: when translating from a low-resource to a high-resource language, LLMs enhanced with retrieval-augmented generation achieve the best results, whereas for high-to-low-resource translation, the fine-tuned `NLLB` model performs better. This conclusion is further supported by a significantly higher proportion of valid words in its generated translations. The incorporation additional data (e.g. through back-translation) would yield better results; however, the primary focus here is on validating the quality of the provided datasets. Overall, these findings underscore both the potential and the necessity of advancing machine translation research for Ladin, as well as the value of the datasets we contribute.

6 Conclusion

In this work, we present our submission to the OLDI shared task 2025, providing FLORES+ translations for Ladin (Val Badia and Gherdëina) and provide an evaluation of several LLMs and the `NLLB` model on this test set, covering six translation directions between Val Badia, Gherdëina, and Italian.

Our results show significant performance gains from the additional parallel datasets released with this work, validating the quality of the datasets and highlighting a promising direction for future research. Beyond fine-tuning neural MT models or relying on basic BM25-Retrieval, our datasets opens the door to more advanced retrieval-augmented prompting strategies, where semantically or syntactically similar sentence pairs are selected to guide translations (Kumar et al., 2023; Merx et al., 2024; Tang et al., 2024; Zebaze et al., 2025).

A central question going forward is whether such carefully designed retrieval and prompting methods could not only provide a more lightweight alternative to fine-tuning for low-resource languages like Ladin, but in some cases even surpass it in translation quality. It would be highly valuable if, in future work, this effort could be extended to the Anpezo, Fassa, and Fodom Ladin variants, enabling their inclusion in the FLORES+ dataset and thus fully representing the Ladin language and accurately reflecting its internal diversity. We hope this work will inspire and encourage further research on Ladin in machine translation, helping to bring the language into sharper focus within the MT community.

Acknowledgements

We would like to thank the anonymous reviewers for their work and valuable comments and suggestions, which have greatly helped to improve our presentation. This initiative was taken as part of the research project *Machine Translation for Ladin* at the University of Innsbruck carried out in collaboration with the Ladin Cultural Institute "Micurá de Rü" and funded by the *Regione Autonoma Trentino–Alto Adige*. We also thank Jürgen Runggaldier for his support in this endeavour.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context Examples Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Roland Bauer. 2022. Language. In Tobia Moroder, editor, *The Ladins of the Dolomites*, pages 28–35. PLUS – University of Salzburg, Salzburg and Bolzano.

Rachel Bawden and François Yvon. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Beatrice Colcuc. 2024. Il ladino. *Linguistik Online*, 130(6):9–30. Creative Commons Attribution 4.0 International License.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and et. al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.

Marco Forni. 2013. *Dizionario italiano – ladino gardenese / Dizioner ladino gardenese – italiano*. Istitut Ladin Micurá de Rü, San Martin de Tor.

Marco Forni. 2019a. *Gramatica Ladin Gherdëina*. Istitut Ladin Micurá de Rü.

Marco Forni. 2019b. *La ortografia dl ladin de Gherdëina*. Istitut Ladin Micurá de Rü, San Martin de Tor.

Samuel Frontull and Georg Moser. 2024a. Rule-based, neural and LLM back-translation: Comparative insights from a variant of Ladin. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 128–138, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Frontull and Georg Moser. 2024b. Traduzione automatica "neurale" per il ladino della Val Badia. *Ladinia*, XLVIII:119–144.

Yuan Gao, Ruili Wang, and Feng Hou. 2024. How to Design Translation Prompts for ChatGPT: An Empirical Study. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia Workshops*, MMAsia '24 Workshops, New York, NY, USA. Association for Computing Machinery.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gabriele Iannaccaro and Vittorio Dell'Aquila. 2004. *La pianificazione linguistica: lingue, società e istituzioni*. Carocci, Roma.

Tom Kocmi and Ondřej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Aswanth Kumar, Ratish Puduppully, Raj Dabre, and Anoop Kunchukuttan. 2023. CTQScorer: Combining Multiple Features for In-context Example Selection for Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7736–7752, Singapore. Association for Computational Linguistics.

Xing Han Lù. 2024. Bm25s: Orders of magnitude faster lexical search via eager sparse scoring.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Giovanni Mischí, Claudia Rubatscher, Isabella Ties, Daria Valentin, and Paul Videsott. 2015. *Grafia nöia - Ladin scrit dla Val Badia: por les scolines y les*

scores ladines. Departimënt Educaziun y Cultura ladina, Balsan.

Sara Moling, Ulrike Frenademetz, and Marlies Valentin. 2016. *Dizionario Italiano–Ladino Val Badia/Dizionar Ladin Val Badia–Talian*. Istitut Ladin Micurá de Rü, San Martin de Tor.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Janine Oliveira, Marison Ranieri Rodrigues de Freitas, Delaney Gomez-Jackson, Juliana Peres Rebelatto Pereira, Natalia Sarmento Tenório Falcão, Roy Yokoyama, Sushil Garg, and Yukitomi Fujinaga. 2024. *Hello Indigenous: a blueprint on the preservation of endangered Indigenous languages through digital inclusion*. Lenovo Foundation and Motorola Mobility, Brazil. UNESCO-sponsored; Electronic version.

Werner Pescosta. 2015. *Storia dei ladini delle Dolomiti*, 2a edizione rielaborata edition. Istitut Ladin Micurà de Rü, San Martin de Tor, Italy.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Chenming Tang, Zhixiang Wang, and Yunfang Wu. 2024. SCOI: Syntax-augmented Coverage-based In-context Example Selection for Machine Translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9956–9971, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Giovanni Valer, Nicolò Penzo, and Jacopo Staiano. 2024. Nesciun Lengaz Lascià Endò: Machine Translation for Fassa Ladin. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 967–975, Pisa, Italy. CEUR Workshop Proceedings.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need.

Paul Videsott. 2023. *Les Ladins des Dolomites*. Peuples en Péril. Éditions Armeline, Crozon.

Paul Videsott, Ruth Videsott, and Jan Casalicchio, editors. 2020. *Manuale di linguistica ladina*, volume 26 of *Manuals of Romance Linguistics*. De Gruyter.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-Context Example Selection via Similarity Search Improves Low-Resource Machine Translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Correcting the Tamazight Portions of FLORES+ and OLDI Seed Datasets

Alp Öktem
Col·lectivaT
alp@collectivat.cat

Mohamed Aymane Farhi
Tamazight NLP
aymenfarhi.25@gmail.com

Brahim Essaidi
Freelance Translator
essaidib2@gmail.com

Naceur Jabouja
Freelance Translator
contact@njtranslations.com

Farida Boudichat
Awal Team
awal@collectivat.cat

Abstract

We present the manual correction of the Tamazight portions of the FLORES+ and OLDI Seed datasets to improve the quality of open machine translation resources for the language. These widely used reference corpora contained numerous issues, including mistranslations, orthographic inconsistencies, overuse of loanwords, and non-standard transliterations. Overall, 36% of FLORES+ and 40% of Seed sentences were corrected by expert linguists, with average token divergence of 19% and 25% among changed items. Evaluation of multiple MT systems, including NLLB models and commercial LLM services, showed consistent gains in automated evaluation metrics when using the corrected data. Fine-tuning NLLB-600M on the revised Seed corpus yielded improvements of +6.05 chrF (en→zgh) and +2.32 (zgh→en), outperforming larger parameter models and LLM providers in en→zgh direction.

1 Introduction

High-quality parallel data is a cornerstone for the development of robust machine translation (MT) systems, particularly for low-resource languages where each sentence can significantly impact model performance. For Tamazight—spoken by over 40 million people across North Africa and the diaspora—parallel corpora are scarce, and widely used datasets such as FLORES+ (Goyal et al., 2022; NLLB Team et al., 2024) and OLDI Seed (Maillard et al., 2023) play a significant role in enabling MT research and evaluation.

The Tamazight portions of FLORES+[1] and OLDI Seed[2] contain Standard Moroccan Tamazight (ISO 639-3: *zgh*), the standardized variety developed by the Royal Institute of Amazigh Culture (IRCAM) (Boukous, 2014) for education and official use since 2001. However, our initial inspection revealed substantial issues: mistranslations, orthographic inconsistencies, malformed or unnecessary loanwords, non-standard transliterations, and occasional semantic inaccuracies. Some of these errors echo broader findings about low-resource language datasets, where insufficient quality control leads to degraded MT outputs and downstream application failures (Kreutzer et al., 2022). Similar issues have been observed for other languages in FLORES+, prompting targeted correction efforts such as those for Hausa, isiZulu, Northern Sotho, and Xitsonga (Abdulmumin et al., 2024). In fact, when FLORES-200 was first released, the Tamazight data was mislabeled as Central Atlas Tamazight (*tzm*) and only corrected following community feedback to its correct code *zgh*.

As part of the Awal project (Öktem and Boudichat, 2025), which develops open-source MT and speech technologies for Tamazight and coordinates community data creation, we undertook a systematic manual correction of the FLORES+ dev and devtest sets (997/1,012 sentences) and the OLDI Seed corpus (6,193 sentences). Corrections were performed by expert linguists using authoritative IRCAM lexicographic and grammatical resources.

The motivation for this work is straightforward: to ensure these widely used benchmark and seed datasets truly reflect Standard Moroccan Tamazight, so that MT systems trained or evaluated on them can produce translations that are both accurate and culturally appropriate. We follow the approach of Abdulmumin et al. (2024) for quantifying the extent of changes. We then evaluate multiple state-of-the-art open-source and commercial MT systems (including LLMs) on both the original and corrected FLORES+ devtest data for English↔Tamazight, and fine-tune existing NLLB

[1] https://huggingface.co/datasets/openlanguagedata/flores_plus

[2] https://huggingface.co/datasets/openlanguagedata/oldi_seed

Proceedings of the Tenth Conference on Machine Translation, pages 1072–1080
November 8-9, 2025

models with the updated Seed data to assess downstream impact.

2 Background

2.1 Tamazight and NLP

Tamazight, also known as Amazigh or Berber[3], is part of the Afro-Asiatic language family and is spoken by over 40 million people across a vast area of North Africa and diaspora communities worldwide (Lafkioui, 2018). It is an official language in Morocco since 2011, Algeria since 2016, Libya since 2017, and Mali since 2023. The standardized form, Standard Moroccan Tamazight (ISO 639-3: *zgh*), was developed by IRCAM from 2001 onwards, drawing on features of the main Moroccan varieties—Tachelhit (*shi*), Tarifit (*rif*), and Central Atlas Tamazight (*tzm*)—and other variants such as Touareg (*tmh*) (Boukous, 2014). As part of this process, IRCAM adopted Tifinaghe-IRCAM (also known as Neo-Tifinagh) in 2003 as the official script, replacing earlier informal use of Latin and Arabic scripts, and providing a phonologically accurate and standardized writing system (Soulaimani, 2016; Ataa-Allah and Boulaknadel, 2012).

The language exhibits considerable dialectal diversity across Morocco, Algeria, and other regions, a reflection of its vast geographic spread. Orthographic variation persists despite standardization, with Latin and Arabic scripts still used informally alongside Neo-Tifinagh. Historical marginalization—shaped by colonization and Arabization—has reduced intergenerational transmission in some areas and within diaspora communities, though Amazigh cultural movements such as the Amazigh Spring played a key role in securing recognition and institutional support (Roque, 2009; CIEMEN and Casa Amaziga de Catalunya, 2019).

From an NLP perspective, Tamazight poses challenges due to its rich morphology, complex orthography, and high dialectal variation. Inflectional and derivational processes, coupled with script diversity, make computational processing for tasks like tokenization, POS tagging, and MT more difficult (Ataa-Allah and Boulaknadel, 2012). Until recently, it remained underrepresented online, with limited user-generated content in the standard form. Notably, the Tamazight Wikipedia was only launched in 2023, marking a significant milestone in its digital presence.

Several important language resources and tools have been released in recent years. IRCAM has developed a Standard Tamazight Corpus (Boulaknadel and Ataa-Allah, 2013), a morphosyntactically annotated corpus (Amri et al., 2017), and tools such as an Amazigh verb conjugator (Ataa-Allah and Boulaknadel, 2014), a concordancer, and a Tifinagh-adapted search engine (Ataa-Allah and Boulaknadel, 2012). More recent research includes Tamazight word embeddings trained on web-collected corpora (Faouzi et al., 2023).

In the MT domain, Tamazight is one of the languages listed within the 200 languages of the *No Language Left Behind (NLLB)* project (NLLB Team et al., 2024), which relies on the FLORES training and evaluation sets (Goyal et al., 2022). Other relevant multilingual datasets that include Tamazight are SIB-200 (Adelani et al., 2024), MADLAD (Kudugunta et al., 2023), and GlotCC/GlotLID (Kargaran et al., 2023, 2025). However, as has been noted in recent research, the development of language technologies for underrepresented languages often follows a top-down approach led by large institutions or technology companies with little input from speaker communities (Moshagen et al., 2024; Bird, 2020; Schwartz, 2022). This dynamic risks misrepresenting languages through technologies built without meaningful community participation, and can perpetuate harm by commodifying indigenous knowledge and sidelining local authorities (Bird, 2020). For marginalized languages, the quality of training data is especially critical: inaccuracies can distort their digital representation and propagate errors through AI systems (Kreutzer et al., 2022; Lau et al., 2025). These issues are not unique to Tamazight—audits of widely used multilingual resources have shown that, for many of the languages they "cover," data quality and representativeness remain poor, creating an illusion of coverage while delivering limited practical usability (Lau et al., 2025).

2.2 Awal initiative

Crowdsourcing initiatives have emerged as a viable alternative to the top-down approaches that dominate language technology development for underrepresented languages. Grassroots efforts such as *Masakhane* (Nekoto et al., 2020), *NaijaVoices*

[3]We include the term "Berber" as it is commonly known in the global north but avoid its usage as it is often considered a pejorative term by the Imazighen (Amazigh people).

(Emezue et al., 2025), and *PARME* (Ahmadi et al., 2025) demonstrate that participatory models can produce data that is both higher quality and better aligned with community norms.

The Awal project (Öktem and Boudichat, 2025) follows this participatory model to create open-source MT and speech resources for Tamazight. Launched in 2024, its main platform, `https://awaldigital.org`, facilitates the translation of sentences from or into Tamazight, covering multiple dialects and scripts. Contributions come from volunteers' own translations or from Creative Commons–licensed material, which can also be post-edited from automatic translations produced by the NLLB engine. The platform integrates a peer-validation feature, where each sentence requires two independent approvals before entering the validated corpus. In addition to text, Awal collects Tamazight speech data through Mozilla Common Voice (Ardila et al., 2020), which is likewise validated via community review. All parallel and monolingual text is openly shared through the project's Hugging Face repository[4], while the voice data is released via Common Voice[5].

Awal's approach combines community-driven data collection with curated dataset creation by professional linguists, as in the present work correcting the Tamazight portions of FLORES+ and OLDI Seed.

3 Methodology

3.1 Correction Workflow

We corrected the Tamazight side of the FLORES+ dev (997 sentences), FLORES+ devtest (1,012 sentences), and OLDI Seed (6,193 sentences) datasets, obtained from the official OLDI Hugging Face repositories.

All sentences were exported into spreadsheets to enable structured, sentence-by-sentence review. The FLORES+ dev and devtest sets were revised in full in two iterations by two linguists. The OLDI Seed dataset was divided into batches of 1,000 sentences and distributed among three professional Tamazight translators, with allocation based on their availability and delivery capacity. To ensure quality control, each linguist's work was spot-checked through random sampling by another linguist.

[4] `https://hf.co/datasets/collectivat/amazic`
[5] `https://commonvoice.mozilla.org`

For the FLORES+ splits, the spreadsheet included the English source sentence, the original Tamazight sentence, and a column for the corrected version. For the OLDI Seed dataset, the spreadsheet additionally included the Arabic translation from the original resource, allowing the translators to use both English and Arabic as context. In all cases, if a sentence required no changes, the original Tamazight sentence was copied into the "corrected" column; if corrections were needed, translators directly post-edited the original, making orthographic, lexical, and syntactic adjustments as appropriate.

Corrections were guided by authoritative lexicographic and grammatical resources from IRCAM and other reference works, including dictionaries (Ameur et al., 2016; Chafik, 1996; Akioud et al., 2022), phonology (Boukous, 2009), and grammar books (Boukhris et al., 2008; El Moujahid, 2022; Laabdelaoui et al., 2012).

3.2 Challenges of Standardization and Dialect Representation

The review process also highlighted the limitations of relying solely on such references for a language that is still undergoing standardisation and must represent multiple dialects. Certain inconsistencies are found even in official resources: for example, masculine nouns are described in the *New Grammar of Amazigh* (Boukhris et al., 2008) as generally beginning with ⴰ, ⵓ, or ⵉ and feminine nouns with ⵜ, yet forms like ⵙⵙⵉⵏⵎⴰ (*ssinima*, 'cinema') or ⵍⴱⴰⵏⴽ (*lbank*, 'bank') appear in IRCAM dictionaries without the expected nominal prefix. Similarly, the rule that a word containing an emphatic consonant should be fully emphatic is not applied consistently across dictionary entries and published texts.

Lexical variation across Tamazight dialects further complicates correction. The verb "to give," for example, appears in IRCAM dictionaries with multiple forms—ⵀⴽ (*fk*), ⴽⵀ (*kf*), ⵓⵛ (*uc*), ⵡⵛ (*wc*), ⵙⵙⵉⵢ (*ssiy*)—without clear indication of which is considered standard. For highly standardised languages, such variation is usually resolved in reference materials, but in Tamazight, the standard form is still being shaped through a combination of institutional policy, community use, and corpus-based practice. This reality means that "correction" work often involves navigating legitimate competing forms rather than simply enforcing a single prescriptive norm.

In addition to these challenges, standardization in Tamazight also involves the ongoing creation of new terms, particularly in scientific and technical domains. This process is not a matter of direct translation but requires applying the language's own morphological strategies for derivation, compounding, and adaptation. As such, parallel data creation and curation go beyond translating sentences: they must be informed by a deep understanding of Tamazight's morphological rules and involve specialists in terminology and linguistic planning. Without this, datasets risk importing external forms rather than contributing to the gradual, community-driven development of a functional standard.

3.3 Error Taxonomy

During the review, we identified and categorized recurrent error types in the original datasets. Below we summarize each type and show representative examples.

3.3.1 Spelling Mistakes. These included violations of standardized Tamazight orthography. Key issues were the improper use of specific characters such as (ⵯ), (ⵯ), and emphatic consonants (ⵅ, ⵇ, ⵜ, ⵚ, ⵒ), which were corrected to align with IRCAM rules. For example, "Open" was corrected from ⵓⵅⵯⵔ to ⵇⵅⵔ, and "Three" from ⴽⵓⴻⵒ to ⴽⵇⴻ. Proper nouns were often transliterated inconsistently, such as "Louis Mayer" which needed to be adjusted from ⵍⵯⵙⵟ ⵔⴰⵣⵟ to ⵍⵍⵙⵟ ⵔⴰⵣⵟ, and "Maria Feodorovna" from ⵔⴰⵟⵙⴰ ⵀⵣⵯⴰⵯⵟⵯⵀⵍⴰ to ⵔⴰⵟⵣⴰ ⵀⵣⵯⴰⵯⵟⵯⵀⵍⴰ. We also found erroneous attachment of pronouns to nouns (e.g. "His film" from ⵍⵀⵙⵍⵎⵟ to ⴰⵀⵙⵍⵔ ⵏⵟ), which were separated for clarity and grammatical correctness.

3.3.2 Transliteration Errors. Particular attention was paid to the accurate transliteration of foreign words and names. The most frequent problems involved the inconsistent mapping of the letter *V* (correctly rendered as ⵀ) and *P* (correctly rendered as ⵜ) in foreign words. For example, "Nova Scotia" was corrected from ⵊⵯⵜⴰ ⵟⴽⵯⵜⵣⴰ to ⵊⵯⵀⴰ ⵟⴽⵯⵜⵣⴰ, and "Sveriges radio" from ⴰⵊⵅⵍⴰⵣ ⵔ ⵟⵜⵙⵟⵣⵠⵟ to ⴰⵊⵅⵍⴰⵣ ⵔ ⵟⵀⵙⵟⵣⵠⵟ. Other non-Tamazight sounds were adjusted to their closest Tamazight equivalents following standardized transliteration practices.

Unnecessary Loanwords		
English	**Original**	**Corrected**
Work	ⵍⵅⴷⵔ	ⵜⴰⵍⵯⵟⵙ
Theatre	ⵍⵔⵟⵇⴰⵗ	ⴰⵔⵅⵠⵯⵊ
Candy	ⵍⵀⵊⵙⴷ	ⵜⴰⵔⵅⵙⴻⵜ
Start	ⵜⴷⵯ	ⵟⵟⵊⵜⵙ
Anemia	ⵀⴰⵣⵟ ⴷⴷⴰⵔⵔ	ⵯⵅⵅⵯⵀ

Malformed Loanwords		
English	**Original**	**Corrected**
Cinema	ⵟⵟⵙⵊⵙⵔⴰ	ⴰⵟⵙⵊⵙⵔⴰ
Film	ⵍⵀⵙⵍⵔ	ⴰⵀⵙⵍⵔ
Television	ⵜⵙⵍⵙⵀⵙⵅⵙⵯⵊ	ⴰⵜⵙⵍⵙⵀⵙⵅⵙⵯⵊ
Oxygen	ⵍⵯⴽⵟⵊⵙⵊ	ⵯⴽⵟⵙⵊⵙⵊ
Nitrogen	ⵏⵙⴻⵇⵯⵊⵙⵊ	ⴰⵊⵙⴻⵇⵯⵊⵙⵊ

Table 1: Unnecessary and malformed loanwords detected in the corpus and their corrections.

3.3.3 Unnecessary Loanwords. Many sentences used Arabic or French loanwords where Tamazight equivalents exist. Following IRCAM's prioritization guidelines, we replaced these with native terms when available, giving preference first to Moroccan Tamazight variants, then to other Tamazight varieties (e.g. Touareg, Kabyle), and retaining loanwords only when unavoidable. Examples are shown in the upper portion of Table 1.

3.3.4 Malformed Loanwords. In cases where loanwords were retained, they often failed to follow Tamazight morphological patterns. Corrections ensured that such words received appropriate prefixes (e.g., ⴰ- for masculine nouns) and were adapted phonologically to fit Tamazight norms. Examples are presented in the lower portion of Table 1.

3.3.5 Mistranslations. Some translations did not accurately reflect the source meaning, introducing semantic drift or mistranslated idiomatic expressions. These were corrected to capture the intended sense and to align with Tamazight syntax. Representative examples are presented in Table 2.

4 Change analysis

We assessed the extent of changes made to the Tamazight portions of the FLORES+ dev/devtest and OLDI Seed datasets using token divergence (Abdulmumin et al., 2024), BLEU (Papineni et al.,

English Source	Original Translation	Back-translation	Correct Translation
Even if you're driving through the subtropical rainforest, a few seconds with the doors open while you get inside the vehicle is enough time for mosquitoes to get in the vehicle with you.	ⵍⵓⵅⵅⵓ ⴰⵓ ⵜⵍⴰⴰⵀⵜ ⵠ ⵜⵓⵅⵓⵍⵉⵍ ⵍ ⵑⵍⵣⵓⵟ ⵉⵕⵍⵓⵍⵍⵓⵍ, ⵓⵕⵕⵠⵕ, ⴰⵕⵑⵕ ⵍ ⵜⵑⵕⴰⵉⴰⵍ ⴰ ⵜⵅⵏⵍⵍⵉ ⵉⵣⵟⵟⵍⵍ ⵜⵉⵏⵉⵍⵍ ⵓⵅⵍⵕⵑ ⵍ ⵜⵉⵕⵕⵑⵑⵑ ⵉⵣⵣⵓ ⵓⵕ ⵉ ⵍⵓⵕⵉⵕⵓ ⵓⴰ ⵉⴰⵅ ⵉⵕⵕⵑⵕ.	Although you are driving in rainy forests like "asbgs", few minutes and the doors open and you are inside the vehicle is fine for mosquitoes to share the vehicle with you.	ⵍⵓⵅⵅⵓ ⴰⵓ ⵜⵍⴰⴰⵀⴰ ⵠ ⵜⵓⵅⵓⵍⵉⵍ ⵍ ⵑⵍⵣⵓⵟ ⵜⵉⵏⵅⵍⵓⵜⵑⵕⵉⵍ, ⵅⵑⵏⵜ ⵅⵕⵕ ⵍ ⵜⵕⵉⵍⵉⵍ ⴰⵕⵑⵕⵍⵉⵍ ⵓⵅⵑⴰ ⵜⵟⵅⵏⵍⵉⵍ ⵓⴰ ⵜⵅⵏⵅⵍⵓⵜⵑⵕⵑ ⵉ ⵍⵓⵕⵉⵕⵓ ⵅⵍⵓ ⵓⴰ ⵉⴰⵅ ⵉⵕⵕⵑⵕ ⵜⵓⵍⵕⵕⵑⵑⵑ.
The walls and roofs of ice caves can collapse and cracks can get closed.	ⵉⵅⵑⴰⵓⵕ ⴰ ⵉⵏⵕⵓⵍ ⵍ ⵑⵅⵕⵓⵉ ⵉⵟⵅⵉ ⴰ ⵉⵜⵜⵑⵉ ⴰ ⵉⵏⵑⵑⵍⵍ ⵓⴰ ⵅⵅⵅⵉⵜ ⵍ ⵉⵏⵕⵓⵍ ⵍ ⵑⵅⵕⵓⵉ ⵓⴰ ⵕⴰⵍⵍ ⴰ ⵅⵕⵜⵜⵉⵍ ⵓⴰ ⵅⵕⵜⵜⵉⵍ.	ⵅⵉⵓⵕⵍ ⵉⵅⵑⴰⵓⵕ ⴰ ⵉⵅⵅⵅⵉⵜ ⵍ ⵉⵏⵕⵓⵍ ⵍ ⵑⵅⵕⵓⵉ ⵓⴰ ⵕⴰⵍⵍ ⴰ ⵉⵕⵜⵜⵉⵍ ⵓⴰ ⵅⵕⵜⵜⵉⵍ.
For a few pennies some children will tell you the story.	ⵏⵓⵕ ⵅⵕⵓ ⵍ ⵜⵕⵓⵕⵑⵏⵉ ⵕⵕⵑⵕⵍ, ⵓⴰ ⵓⵅ ⵓⵑⵕⵍ ⵉⵇⵉⵕⵕⵓⵍ ⵜⵓⵍⵑⵟⵕⵜ.	Just some few steps children will tell you the story.	ⵏⵓⵕ ⵅⵕⵓ ⵍ ⵜⵓⵕⵕⵓⵑⵑⵉ ⵕⵕⵑⵕⵍⵉⵍ, ⵕⵓⴰ ⵓⵅ ⵓⵑⵕⵍ ⵉⵇⵉⵕⵕⵓⵍ ⵜⵓⵍⵑⵟⵕⵜ.

Table 2: Examples of mistranslations and their corrections. For each case, we show the original English source sentence, the problematic Tamazight translation from the dataset, a back-translation of that Tamazight into English, and the corrected Tamazight translation.

dataset	#rows	#corr. (%)	% token div.	$BLEU_c$	TER_c	WER_c	CER_c
dev	997	384 (39%)	19.36	81.06	12.22	13.35	5.70
devtest	1012	339 (34%)	19.46	79.85	12.35	13.50	5.48
FLORES+	2009	723 (36%)	19.41	78.26	14.84	17.73	8.83
seed	6193	2490 (40%)	25.01	80.49	12.28	13.42	5.60
ALL	8202	3213 (39%)	23.75	78.72	14.29	16.76	8.10

Table 3: Correction statistics and edit-distance metrics for each dataset. Shows the number of sentences corrected (as percentage of total) and average change metrics computed only on modified sentences. FLORES+ aggregates dev and devtest splits, while COMBINED combines FLORES+ and OLDI Seed datasets.

dataset	>50% (%)	>80% (%)
dev	20 (5.2%)	6 (1.6%)
devtest	26 (7.7%)	3 (0.9%)
seed	316 (12.7%)	62 (2.5%)
FLORES+	46 (6.4%)	9 (1.2%)
ALL	362 (11.3%)	71 (2.2%)

Table 4: Number and percentage of corrected sentences with token divergence greater than 50% and 80%, indicating substantial rewrites.

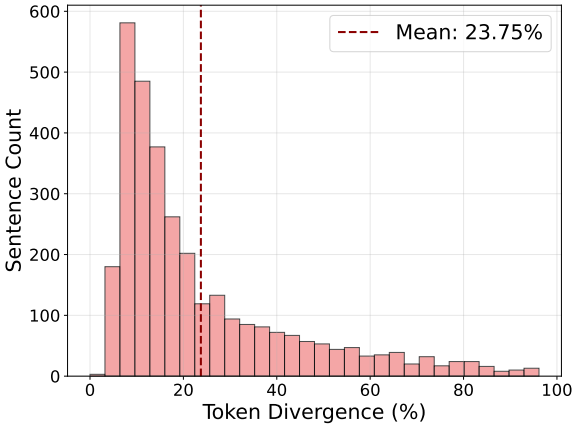


Figure 1: Token divergence distribution across all corrected sentences in the combined FLORES+ and OLDI Seed datasets.

2002), Translation Edit Rate (TER) (Snover et al., 2006), Word Error Rate (WER), and Character Error Rate (CER). Following Abdulmumin et al. (2024), all metrics were computed only on sentences that were modified, avoiding dilution by unchanged items.

Token divergence was calculated as:

$$\text{divergence} = \frac{|T_o - T_c| + |T_c - T_o|}{|T_o \cup T_c|} \quad (1)$$

where T_o is the set of tokens in the original sentence and T_c is the set of tokens in the corrected sentence. This metric measures the proportion of unique tokens that differ between the two versions of a sentence, with 0 indicating identical token sets and 1 indicating no overlap. Reported values are averages over all changed sentences and expressed as percentages.

Table 3 reports, for each dataset, the proportion of sentences corrected and the mean metric scores. In addition to individual datasets, we also report aggregated results for FLORES+ overall

(dev+devtest) and for the combined FLORES+ and Seed datasets. Figure 1 presents the token divergence distribution across all corrected sentences in the combined dataset.

The results show that the proportion of sentences corrected is 36% for FLORES+ overall and 40% for the Seed dataset. Average token divergence values are 19% for FLORES and 25% for Seed, indicating that corrections typically involved localised changes to a minority of tokens in a sentence rather than complete rewrites. This is consistent with the high BLEU scores (≈78–81) and the relatively modest WER (13–18%) and CER (5–8%) values. While substantial rewrites are relatively rare, they are still present in both datasets. Table 4 shows the proportion of corrected sentences whose token divergence exceeded 50% and 80%. For example, 6.4% of corrected sentences in FLORES overall and 12.7% in the Seed dataset had over 50% divergence, indicating a reworking of more than half of their token content.

5 Machine Translation Evaluation

We evaluated English↔Tamazight (zgh) translation quality on the FLORES+ devtest split, comparing results obtained with the original and corrected versions of the dataset. The evaluation covered:

- **Baseline models**: NLLB checkpoints 600M, 1.3B, 1.3B Distilled, 3.3B
- **Fine-tuned models**:
  - i) NLLB-600M fine-tuned on the corrected OLDI Seed dataset for 3 epochs.
  - ii) NLLB-600M fine-tuned for 1.25 epochs on the corrected OLDI Seed dataset and other parallel data of approximately 50,000 segments sourced from Awal, Tatoeba[6], and Tamazight NLP[7] initiatives.
- **Commercial LLMs**: Gemini Pro 2.5, Claude 3.5, and Claude 3.7[8].

Scores were computed using BLEU, chrF++ (Popović, 2015), and TER (Snover et al., 2006) for both translation directions. Table 5 reports the results, with each cell showing the original dataset score / corrected dataset score format.

Across all models tested, corrections led to consistent improvements in BLEU and chrF++, with corresponding reductions in TER. We see an average chrF increase of 0.14 points in en→zgh, and 0.23 in zgh→en direction among all off-the-shelf models and LLM services. This confirms that noise in the original FLORES+ Tamazight references measurably affected evaluation quality.

Fine-tuning on the corrected OLDI Seed dataset (NLLB-0.6B-FT) yielded substantial gains compared to the same 0.6B baseline: +6.05 chrF and +3.05 BLEU in the en→zgh direction. The fine-tuned model achieved 31.69 chrF, performing better than all other larger models and LLM providers, showing that carefully curated data can yield substantial improvements.

In the zgh→en direction, the fine-tuned 0.6B model improved with +2.32 chrF and +1.89 BLEU from corrections, however it was still outperformed by other models, with Gemini Pro 2.5 achieving the highest chrF score of 44.29.

We also report results with additional parallel data (NLLB-0.6B-FT+): 32.71 chrF and 8.84 BLEU in en→zgh, and 40.66 chrF and 18.29 BLEU in zgh→en.

6 Conclusions

We carried out a systematic manual correction of the Tamazight portions of the FLORES+ dev/devtest and OLDI Seed datasets, resolving orthographic inconsistencies, mistranslations, and problematic loanwords. These corrections, performed by professional linguists with reference to authoritative resources, affected a substantial share of sentences and resulted in more accurate benchmarks for MT evaluation.

Across all off-the-shelf models and commercial LLMs tested, the corrected data yielded consistent improvements in automatic evaluation metrics. Fine-tuning on the corrected OLDI Seed dataset further demonstrated the impact of these revisions: the NLLB-600M model trained on the revised data outperformed larger parameter models and LLM providers in en→zgh direction, and recorded improvements in zgh→en direction.

Beyond the experiments, our work underlines broader challenges of dataset creation for a language still undergoing standardization, where legit-

[6] https://tatoeba.org

[7] https://huggingface.co/Tamazight-NLP

[8] We also intended to evaluate Google Translate; however, its API does not currently support Tamazight, making programmatic evaluation impractical, so it was excluded from this study.

Model	BLEU ↑		chrF++ ↑		TER ↓	
	en-zgh	zgh-en	en-zgh	zgh-en	en-zgh	zgh-en
NLLB-0.6B	4.86/4.96	14.84/15.3	25.52/25.64	37.1/37.38	94.97/94.77	77.27/76.06
NLLB-0.6B-FT	7.8/8.01	17.13/17.19	31.46/31.69	39.51/39.7	83.72/83.32	73/73.01
NLLB-0.6B-FT+	8.64/8.84	17.97/18.29	32.5/32.71	40.27/40.66	82.17/81.87	71.72/71.18
NLLB-0.6B-FT+, S=2	9.12/9.38	18.29/18.58	33.26/33.52	40.45/40.87	81.38/**80.97**	70.71/70.45
NLLB-0.6B-FT+, S=8	9.42/**9.69**	18.37/18.72	33.79/**34.05**	40.55/40.91	81.92/81.49	70.63/70.38
NLLB-1.3B Dist.	6.84/7.02	18.16/18.32	28.93/29.07	39.73/39.97	86.24/85.99	72.05/71.68
NLLB-1.3B	6.34/6.37	17.23/17.35	27.35/27.44	39.11/39.33	89.75/89.59	73.28/72.85
NLLB-3.3B	7.52/7.68	17.72/18.05	29.7/29.8	39.43/39.8	83.52/83.42	72.81/72.14
Gemini Pro 2.5	4.7/4.83	21.34/**21.31**	26.4/26.56	44.35/**44.29**	86.59/86.27	69.16/**69.23**
Claude 3.5	5.43/5.51	17.41/17.65	27.33/27.42	38.52/38.81	82.25/82.19	74.62/74.24
Claude 3.7	5.47/5.51	17.95/18.09	28.1/28.18	41.29/41.43	83.31/83.17	73.5/73.13

Table 5: Machine translation evaluation results on FLORES+ devtest for English↔Tamazight translation. Each cell shows original dataset score / corrected dataset score. NLLB-0.6B-FT refers to the model fine-tuned on the corrected OLDI Seed dataset, while NLLB-0.6B-FT+ includes additional parallel data. S refers to the beam size used when decoding using beam search; when not specified, greedy search is used. Bold values mark the best performing metric on corrected evaluation.

imate variation coexists with quality issues. It also illustrates how massively parallel datasets can be problematic for low-resource languages: without careful linguistic validation, they risk amplifying errors and misrepresenting the language in downstream systems.

Future work will extend fine-tuning experiments with additional parallel data and explore other language pairs involving Tamazight, ensuring that corrected datasets serve as a stronger foundation for both evaluation and system development.

Acknowledgments

This work was conducted as part of the *Som Part* project, led by CIEMEN and the Fundació pels Drets Col·lectius dels Pobles, with funding from the Catalan Agency for Development Cooperation (ACCD) and the Municipality of Barcelona.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

David Adelani, Hong Liu, Xu Shen, Nikita Vassilyev, Jesujoba Alabi, Yuxiang Mao, Hongyu Gao, and Eric Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245. Association for Computational Linguistics.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, and 14 others. 2025. PARME: Parallel corpora for low-resourced Middle Eastern languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30032–30053, Vienna, Austria. Association for Computational Linguistics.

Hassan Akioud, Meftaha Ameur, Khalid Ansar, Abdessalam Boumasser, and Noura El Azrak. 2022. *Arabic-Amazigh-French Landforms Dictionary*. Royal Institute of Amazigh Culture.

Meftaha Ameur, Khalid Ansar, Abdellah Boumalk, Noura El Azrak, Rachid Laabdelaoui, and Hamid Souifi. 2016. *Dictionnaire Général de la Langue Amazighe*. Institut Royal de la Culture Amazighe.

Samira Amri, Lahcen Zenkouar, and Mustapha Outahajala. 2017. Build a morphosyntaxically annotated amazigh corpus. In *Proceedings of the 2nd International Conference on Big Data, Cloud and Applications*. Association for Computing Machinery.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Fadoua Ataa-Allah and Siham Boulaknadel. 2014. Amazigh verb conjugator. In *Proceedings of the*

Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 1051–1055, Reykjavik, Iceland. European Language Resources Association (ELRA).

Fouad Ataa-Allah and Sada Boulaknadel. 2012. Toward computational processing of less resourced languages: Primarily experiments for moroccan amazigh language. In S. Sakurai, editor, *Theory and Applications for Advanced Text Mining*, chapter 9. IntechOpen.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fatima Boukhris, Abdallah Boumalk, El Houssaïn El Moujahid, and Hamid Souifi. 2008. *La Nouvelle Grammaire de l'Amazighe*. Institut Royal de la Culture Amazighe.

Ahmed Boukous. 2009. *Phonologie de l'Amazighe*. Institut Royal de la Culture Amazighe.

Ahmed Boukous. 2014. The planning of standardizing amazigh language: The moroccan experience. *Îles d'Imesli*, 6(1):7–23.

Sada Boulaknadel and Fouad Ataa-Allah. 2013. Building a standard amazigh corpus. In *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011)*, pages 91–98, Prague, Czech Republic.

Mohammed Chafik. 1996. *Arabic-Amazigh Dictionary*. Academy of the Kingdom of Morocco.

CIEMEN and Casa Amaziga de Catalunya. 2019. *El poble amazic a Catalunya*. CIEMEN.

ElHoussain El Moujahid. 2022. *Grammaire Générative de l'Amazighe: Morphologie et Syntaxe du Nom*. Institut Royal de la Culture Amazighe.

Chris Emezue, NaijaVoices Community, B. Awobade, A. Owodunni, H. Emezue, G. M. T. Emezue, N. N. Emezue, S. Ogun, B. Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. In *Proceedings of Interspeech 2025*, pages August 17–21, Rotterdam, Netherlands. International Speech Communication Association (ISCA).

Hicham Faouzi, Mohammed El-Badaoui, Mohammed Boutalline, Abdelaziz Tannouche, and Hamza Ouanan. 2023. Towards amazigh word embedding: Corpus creation and word2vec models evaluations. *Revue d'Intelligence Artificielle*, 37(3):753–759.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amirhossein Kargaran, Amirhossein Imani, François Yvon, and Hinrich Schuetze. 2023. Glotlid: Language identification for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4863–4875. Association for Computational Linguistics.

Amirhossein H. Kargaran, François Yvon, and Hinrich Schütze. 2025. Glotcc: an open broad-coverage commoncrawl corpus and pipeline for minority languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pages 16983–17005, Red Hook, NY, USA. Curran Associates Inc.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Rachid Laabdelaoui, Abdallah Boumalk, El Mehdi Iazzi, Hamid Souifi, and Khalid Ansar. 2012. *Manuel de Conjugaison de l'Amazighe*. Institut Royal de la Culture Amazighe.

Mena B. Lafkioui. 2018. *Berber Languages and Linguistics*. Oxford Bibliographies.

Michael Lau, Qi Chen, Yuchen Fang, Tian Xu, Tong Chen, and Pavel Golik. 2025. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Svein N. Moshagen, Lene Antonsen, Linda Wiechetek, and Trond Trosterud. 2024. Indigenous language technology in the age of machine learning. *Acta Borealia*, 41(2):102–116.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Alp Öktem and Farida Boudichat. 2025. Awal – community-powered language technology for tamazight. In *Proceedings of the Conférence Internationale sur les Technologies d'Information et de Communication pour l'Amazighe (TICAM)*, Rabat, Morocco. Institut Royal de la Culture Amazighe (IRCAM). Submitted.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. Chrf: Character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon.

Mikel Roque. 2009. *Els amazigs avui, la cultura berber*. Pagès Editors / IEMed.

Lane Schwartz. 2022. Primum non nocere: Before working with indigenous data, the acl must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 724–731, Dublin, Ireland. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, USA.

Driss Soulaimani. 2016. Writing and rewriting amazigh/berber identity: Orthographies and language ideologies. *Writing Systems Research*, 8(1):1–16.

Filling the Gap for Uzbek: Creating Translation Resources for Southern Uzbek

Mukhammadsaid Mamasaidov[1] **Azizullah Aral**[2]
Abror Shopulatov[1,3] **Mironshoh Inomjonov**[1]
[1] Tilmoch [2] Institute of Uzbek National Musical Art [3] MBZUAI

Abstract

Southern Uzbek (uzs) is a Turkic language variety spoken by around 5 million people in Afghanistan and differs significantly from Northern Uzbek (uzn) in phonology, lexicon, and orthography. Despite the large number of speakers, Southern Uzbek is underrepresented in natural language processing. We present new resources for Southern Uzbek machine translation, including a 997-sentence FLORES+ dev set, 39,994 parallel sentences from dictionary, literary, and web sources, and a fine-tuned NLLB-200 model (lutfiy). We also propose a post-processing method for restoring Arabic-script half-space characters, which improves handling of morphological boundaries. All datasets, models, and tools are released publicly to support future work on Southern Uzbek and other low-resource languages.

1 Introduction

The Southern Uzbek language, spoken by approximately 5 million Uzbeks residing across 14 provinces of Afghanistan, represents a distinct linguistic variety that has developed independently from Northern Uzbek over centuries (Ethnologue, 2025a). Uzbek as a whole is classified as a macrolanguage according to ISO 639-3 standards, encompassing multiple related varieties including Northern Uzbek (uzn) spoken primarily in Uzbekistan, and Southern Uzbek (uzs) prevalent in Afghanistan (Ethnologue, 2025b).

This macrolanguage classification recognizes the significant linguistic diversity within the broader Uzbek language family, where individual varieties have developed distinct phonological, lexical, and grammatical features due to geographical separation and contact with other languages. As part of the global Uzbek population exceeding 34 million people, Southern Uzbek is recognized in Afghanistan's Constitution as a potential third official language in regions where it is the majority language, in addition to Pashto and Dari. (Afghanistan, 2004)

Southern Uzbek functions as a fully developed literary language that meets the demands of literature, art, culture, and science. It maintains active presence across multiple domains including technology, education, diplomacy, banking, and commerce. The language is taught in Southern Uzbek departments at seven national universities in Afghanistan and serves as the medium of instruction in 970 schools distributed across provinces: 9 schools in Badakhshan, 80 in Balkh, 450 in Faryab, 50 in Samangan, 300 in Sar-e-Pol, and 80 in Takhar. (Olim Labib, 2020)

International media outlets including BBC, Radio Free Europe/Radio Liberty (Ozodlik), Voice of America, Voice of Iran, TRT Avaz, and Sputnik actively broadcast in Southern Uzbek, alongside Afghan media channels such as Oyna, Botur, Almas, Orzu, Nur, Oriano, Kalid, and National Radio and Television. The language maintains expanding digital presence across major online platforms including Wikipedia, Google, Facebook, and other social networks.

Despite this linguistic vitality, Southern Uzbek remains underrepresented in natural language processing technologies. Major translation platforms like Google Translate (Google, 2025) currently provide limited or no support for this language variety, highlighting the critical need for dedicated computational resources. As a low-resource language with unique characteristics distinct from Northern Uzbek, Southern Uzbek presents significant challenges for machine translation systems.

This study, conducted as part of the Open Language Data Initiative (OLDI) shared task, addresses these challenges by developing specialized neural machine translation models for Southern Uzbek. Our contributions parallel recent advances in low-resource language processing and include:

1. A FLORES+ dev dataset translated to South-

Proceedings of the Tenth Conference on Machine Translation, pages 1081–1087
November 8-9, 2025

ern Uzbek containing 997 sentences

2. Parallel corpora for various language pairs with Southern Uzbek

3. Open-sourced fine-tuned neural models for Southern Uzbek translation

4. Comprehensive evaluation against existing baselines

Our research aims to advance machine translation capabilities for Southern Uzbek, contributing to the larger OLDI objective of expanding linguistic diversity in NLP technologies for underrepresented language varieties.

2 Linguistic Background

2.1 Historical Development

Southern Uzbek belongs to the Turkic language family, specifically derived from the Karluk-Chigil-Uyghur dialectal group with partial influences from the Kipchak and Oghuz branches. The language represents the contemporary form of a literary tradition spanning over a millennium, with historical continuity traceable through classical poets including Khwarizmi, Lutfi, Atayi, Sakkaki, Navoi, Babur, lutfiy, and Ogahi. Notably, while these historical figures did not identify themselves as "Uzbek", they wrote in a language that forms the foundation of modern Southern Uzbek, demonstrating the language's independent development into a mature linguistic system. (Habibi Aral, 2021)

Historically, Southern Uzbek served as the administrative and literary language for major dynasties including the Yaftids, Kushans, Ghaznavids, Seljuks, Timurids, and Mughals, who governed territories across Afghanistan and India for centuries using this language and established profound cultural legacies. (Tursunov and O'rinboyev, 1982)

2.2 Writing System

Southern Uzbek employs the Arabic script, which has served as the official writing system for Afghan languages for over a thousand years. This orthographic system presents unique challenges and characteristics that distinguish it from Latin-based Northern Uzbek.

The Arabic-based script includes only three vowel letters: ا (a/o), و (u/o'), and ی (i/y). This limited vowel representation often misleads learners into believing that Uzbek contains only three vowel sounds. However, vowel quality distinctions become evident in minimal pairs such as shown in Figure 1.

kuz (autumn) (کوز)	ko'z (eye) (کۆز)
yel (wind) (یېل)	yil (year) (ییل)
qurol (weapon) (قورال)	maral (deer) (مرەل)

Figure 1: Vowel differences in Southern Uzbek

Standard Uzbek contains six primary vowels (with additional dialectal variants), yet the Arabic script lacks direct representation for half of them. These vowels require indication through diacritical marks (fatha, damma, kasra), which are frequently omitted in practical writing, thereby complicating accurate reading and pronunciation.

Additional complexity arises from the dual functionality of certain letters. The Arabic letter ه (h) functions both as vowel and consonant. Similarly, letters و and ی (waw and ya) serve dual roles as vowels and consonants ("v" and "y") depending on context as illustrated in Figure 2.

Uzbek Southern	Uzbek Northern	Sound
Letter ه (h) dual roles		
بیلدیرەدی	bildiradi	/a/
هوس	havas	/h/
Letter و (waw) dual roles		
وطن	vatan	/v/
توز	tuz	/u/
Letter ی (ya) dual roles		
بای	boy	/y/
فیل	fil	/i/

Figure 2: Dual letters in Southern Uzbek

Southern Uzbek	Northern Uzbek	Meaning
Examples with suffix "-chi"		
چایخانه‌چی	choyxonachi	teahouse keeper
ادبیاتچی	adabiyotchi	writer
Compound words and prefixes		
بي‌تشویش	betashvish	carefree
نا‌انصاف	noinsof	dishonest

Figure 3: Examples of standardized Southern Uzbek Arabic-script orthography showing mandatory half-space (zero-width non-joiner, U+200C) placement. Red marks indicate the location of half-spaces in suffixation after vowel-final stems and in prefix attachment.

Arabic and Persian loanwords maintain their original orthographic forms, typically without vowel markings.

2.3 Morphological Structure

Southern Uzbek exhibits rich agglutinative morphology characteristic of Turkic languages. The language employs extensive suffixation systems that can be classified into various functional categories:

- Nominalizers (noun-forming suffixes)
- Adjectival suffixes
- Verb formers
- Tense and aspect markers
- Other functional and derivational affixes

Standardized orthographic rules govern affix attachment in Southern Uzbek Arabic script. A fundamental principle distinguishes between suffixes attached to vowel-final versus consonant-final stems (-chi, -chilik, -lik, -li, etc.).

These suffixes require half-space (also known as zero-width non-joiner, U+200C, also found in Farsi) separation when attached to stems ending in vowels (represented by Arabic letters ه, و, ا), while connecting directly to consonant-final stems.

Southern Uzbek also employs prefixes, commonly found in Persian or Arabic loanwords, for forming adjectives or adverbs. These prefixes (be-, no-, xo‘sh-, ser-, ba-, ham-, bad-) are written with half-space separation, as shown in Figure 3.

2.4 Contemporary Status and Challenges

Despite its historical significance, Southern Uzbek has faced political marginalization over the past three centuries, with Turkic peoples in Afghanistan being sidelined in governance and education. Progress began in the 1970s when Uzbek parliamentary representatives secured broadcasting rights on Afghan national radio. The 1978 rise of the People’s Democratic Party marked further advancement with the publication of the Yulduz newspaper in Southern Uzbek, establishment of Uzbek Language and Literature departments, and expansion of Uzbek-medium education. (Aral, 2025)

The 2001 democratic reforms in Afghanistan formally granted Southern Uzbek official status, recognizing its role in Afghan multilingual society. However, challenges remain in standardizing orthographic practices and developing computational resources for this linguistically rich but technologically underrepresented variety.

3 Related Work

Machine translation for low-resource languages has gained significant attention, with researchers exploring various approaches from data augmentation to multilingual transfer learning. Dale (2022) developed the first neural MT system for Erzya, a low-resource Uralic language, demonstrating how extensive data mining from diverse sources (Bible texts, dictionaries, digitized books) can yield functional translation systems despite limited parallel data. Similarly, P M et al. (2024) focused on low-resource Indic languages by fine-tuning multilingual models and employing back-translation with careful quality filtering, showing that selective data augmentation can improve performance when synthetic data is judiciously filtered.

Goyle et al. (2023) systematically evaluated strategies for compensating data scarcity in languages like Sinhala, Nepali, Khmer, and Pashto. They found that combining back-translation with focal loss yields substantial improvements, particularly when leveraging large monolingual corpora and transfer learning from related high-resource languages.

Recent advances in large language models have also shown promise for low-resource translation tasks. Commercial LLMs like GPT-4 and Claude demonstrate multilingual capabilities that extend to languages not explicitly included in their training data, offering competitive performance through few-shot learning approaches.

Despite these advances, Southern Uzbek remains largely unexplored in computational linguistics. While Northern Uzbek has received some attention in multilingual models like NLLB (NLLB Team et al., 2022) and MADLAD-400 (Kudugunta et al., 2024), the Southern Uzbek has been left behind. Our work represents the first dedicated effort to develop neural translation resources for this variety of Uzbek.

4 Datasets

4.1 FLORES+ Dev Dataset

This study introduces the Southern Uzbek FLORES+ dev dataset, comprising 997 sentences translated from English to Southern Uzbek (see Figure 4).

The dataset was developed under the Open Language Data Initiative (OLDI) framework. One native Southern Uzbek linguist was responsible

English	The aircraft had been headed to Irkutsk and was being operated by interior troops.
Northern Uzbek	Samolyot Irkutsk tomon yo‘l olgan va ichki qo‘shinlar tomonidan boshqarilayotgan edi.
Southern Uzbek	اوچاق ایرکوتسک تامان یۏل آلگن و ایچکی قۏشینلر تامانیدن باشقەریلەیاتگن اېدی.

Figure 4: Example from the FLORES+ dataset in English, Northern Uzbek and Southern Uzbek.

for the translation process, with subsequent post-review process to ensure linguistic accuracy and cultural appropriateness. All Southern Uzbek translations strictly adhere to the Arabic script orthographic conventions, including proper implementation of half-space characters (U+200C) for morphological boundaries as described in Section 2.3.

Given the complexity of Arabic script representation and the morphologically rich nature of Southern Uzbek, particular attention was paid to maintaining consistent orthographic standards throughout the translation process. The translation process followed standardized conventions for affix attachment, vowel representation, and proper handling of Arabic and Persian loanwords within the Southern Uzbek linguistic system.

4.2 Training Data

The training dataset comprises diverse parallel corpora sourced from three primary domains, totaling 39,994 sentence pairs across multiple language combinations:

1. **Dictionary Entries (1,550 pairs)**: Parallel dictionary entries mapping Northern Uzbek to Southern Uzbek lexical items (Aral, 2024). These entries provide direct lexical correspondences and serve as high-quality alignment data for closely related language varieties.

2. **Literary Corpus (35,865 pairs)**: Parallel sentences extracted through careful alignment from 27 selected books available in both Northern and Southern Uzbek variants. This corpus represents the largest component of our training data and captures literary register variations, complex syntactic structures, and cultural terminology.

3. **Web-sourced Content (2,579 pairs)**: Parallel sentences of English-Southern Uzbek mined from official government websites and reliable online resources. This component provides contemporary usage patterns and domain-specific terminology from governmental and institutional contexts.

4.3 Data Mining Process

The sentence alignment process presented unique challenges due to Southern Uzbek's underrepresentation in existing multilingual models. Our alignment methodology employed a two-stage approach to maximize extraction efficiency.

For literary corpus alignment, we initially applied LaBSE embeddings (Feng et al., 2020) directly to the original Arabic script texts. While LaBSE does not include Southern Uzbek in its training data, the model demonstrated limited alignment capability, likely due to shared vocabulary with other Turkic languages in the embedding space.

To improve alignment quality, we implemented a transliteration-based enhancement strategy. Southern Uzbek texts were transliterated from Arabic to Latin script using rule-based conversion scripts[1], which enabled more effective cross-lingual embedding alignment. This transliteration approach yielded a 40% more successfully aligned sentence pairs compared to direct Arabic script processing.

The sentence alignment methodology follows established practices from low-resource language processing (Dale, 2022). We utilize LaBSE to generate embeddings for each potential sentence pair, calculate cosine similarity between embeddings, and adjust similarity scores using length ratios.

For web-sourced English-Southern Uzbek data, we employed a reverse translation verification approach. Southern Uzbek sentences were translated to English using Gemini-2.0-Flash, followed by LaBSE-based alignment between original English content and back-translated English. This process underwent manual review to ensure translation quality and semantic fidelity.

A notable preprocessing challenge emerged regarding half-space character consistency. Due to OCR limitations and editorial inconsistencies in source materials, half-space characters (U+200C) were frequently omitted, incorrectly rendered as full spaces, or merged with adjacent characters. While this issue complicates training data quality, we address it through post-processing correction

[1] https://github.com/tahrirchi/uzs-scripts

Model	uzs-en	uzs-uzn	eng-uzs	uzn-uzs
gpt-4.1	24.90 / 53.42	2.634 / 3.657	0.48 / 9.49	1.42 / 21.55
gemini-2.0-flash-001	**32.81 / 58.80**	**62.45** / 73.67	**1.59** / 24.47	6.96 / 41.11
claude-sonnet-4	22.25 / 51.46	59.18 / **83.63**	0.68 / 15.38	2.62 / 28.85
nllb-200-600M	3.73 / 23.88	4.14 / 27.02	-	-
Google Translate	9.56 / 33.58	5.13 / 33.19	-	-
madlad400-3b-mt	2.95 / 23.26	0.19 / 1.41	-	-
lutfiy (no half-space fix)	11.26 / 34.39	53.48 / 78.54	1.33 / 25.43	25.99 / 66.44
lutfiy (with half-space fix)			1.58 / **26.61**	**34.31 / 71.11**

Table 1: Evaluation of several models on sacreBLEU/chrF++ across various language pairs involving English, Northern Uzbek (uzn) and Southern Uzbek (uzs).

mechanisms described in Section 4.

5 Translation Experiments

5.1 Model Training

Our experimental framework employed the nllb-200-distilled-600M model as the foundation for Southern Uzbek machine translation development. e maintained the original tokenizer configuration, leveraging the model's existing multilingual capabilities for Turkic language processing.

5.1.1 Training Configuration

For the training process we employed the Adafactor (Shazeer and Stern, 2018) optimizer paired with a learning rate of 1×10^{-4} following a constant schedule and 1000 warmup steps. A weight decay of 1×10^{-3} was applied, and the batch size was set to 32 due to GPU memory constraints. The maximum sequence length was limited to 128 tokens, and training was conducted for 5000 steps, corresponding to approximately 2–3 epochs. All experiments were run on a single A100 40GB GPU. The Adafactor optimizer was chosen for its memory efficiency and proven effectiveness in transformer fine-tuning scenarios, while the conservative learning rate and weight decay values were selected to mitigate overfitting given the small size of the training dataset.

5.1.2 Model Variant

We fine-tuned nllb-200-distilled-600M (NLLB Team et al., 2022) model on the complete 39,994 sentence pair corpus. Our model called **lutfiy**[2] maintains the original NLLB tokenizer and vocabulary, relying on existing Turkic language representations for Southern Uzbek processing.

5.1.3 Half-Space Post-Processing

A critical technical challenge emerged regarding the handling of half-space characters. The NLLB SentencePiece (Kudo and Richardson, 2018) tokenizer normalizes half-space characters (U+200C) to regular spaces during preprocessing, preventing the model from learning proper morphological boundary representation. This problem affects not only Southern Uzbek but also extends to other languages requiring half-space characters, including Persian (Doostmohammadi et al., 2020).

To address this limitation, we developed a character-level n-gram post-processing model that predicts half-space insertion positions. The model was trained on a small set of training data with corrected half-space characters. It analyzes character sequences and applies statistical rules to determine whether half-spaces should follow specific vowel endings in morphologically complex constructions.

This approach provides a practical solution to the tokenizer normalization problem while maintaining compatibility with existing NLLB infrastructure. The post-processing correction mechanism is made publicly available alongside our trained models[3].

5.2 Evaluation Framework

Model performance was assessed using two widely adopted metrics for translation tasks: **sacreBLEU** (Post, 2018), a standardized BLEU implementation that ensures consistent n-gram precision measurement across experiments, and **chrF++** (Popović, 2017), a character-level F-score metric that is particularly well-suited for evaluating morphologically rich languages such as Southern Uzbek. All results are reported on the FLORES+ dev set, enabling comparability with other low-resource language initiatives under the OLDI framework.

[2]Lutfi, a 15th-century Central Asian poet

[3]https://huggingface.co/tahrirchi/lutfiy

6 Results and Discussion

Our evaluation on the FLORES+ Southern Uzbek dev set reveals several key insights into the performance of various translation approaches. The results, presented in Table 1, demonstrate significant performance variations across different model architectures and translation directions.

Notably, large language models exhibit superior performance in understanding Southern Uzbek content, particularly in **uzs-*** directions. Gemini-2.0-Flash achieves the highest scores for uzs-en translation (32.81 BLEU/58.80 chrF++), while Claude-Sonnet-4 excels in uzs-uzn translation quality (83.63 chrF++). This suggests that LLMs' extensive multilingual pretraining enables effective comprehension of low-resource language varieties, even without explicit training on Southern Uzbek data. In contrast, traditional MT systems like Google Translate and specialized multilingual models (NLLB-200-600M, MaLLaD400) demonstrate substantially lower performance, highlighting the challenges these architectures face with underrepresented languages.

However, our fine-tuned **lutfiy** model demonstrates clear advantages in generation tasks. For translation **into** Southern Uzbek (uzn-uzs), our model consistently outperforms all baselines, achieving 34.31 BLEU/71.11 chrF++ for uzn-uzs directions. This validates our approach of fine-tuning on domain-specific parallel corpora, as the model learns proper Southern Uzbek generation patterns that generic LLMs cannot replicate effectively.

The impact of our half-space post-processing correction is particularly evident in the uzn-uzs translation pair. While chrF++ scores show modest improvements (from 66.44 to 71.11), BLEU scores increase dramatically (from 25.99 to 34.31), representing a 32% relative improvement. This substantial BLEU gain with stable chrF++ performance indicates that the half-space correction primarily addresses tokenization boundary issues rather than fundamental translation errors. Since BLEU relies on exact n-gram matches, incorrect half-space placement can artificially deflate scores even when the underlying translation quality remains high.

For the closely related uzs-uzn translation direction, Gemini-2.0-Flash demonstrates exceptional generation capability (62.45 BLEU), significantly outperforming our specialized model (53.48 BLEU). This suggests that LLMs may be particularly effective at cross-dialectal translation within the same language family, possibly due to their ability to capture subtle linguistic variations during pretraining.

These findings highlight complementary strengths between LLMs and specialized fine-tuned models: while LLMs excel at understanding and translating from Southern Uzbek, targeted fine-tuning proves essential for high-quality generation into Southern Uzbek, particularly for morphologically complex constructions requiring proper orthographic conventions.

7 Conclusion

Our study presents the first comprehensive neural machine translation resources for Southern Uzbek, addressing a significant gap in computational linguistics for this underrepresented Turkic variety. Our key contributions include:

1. Creation of a 997-sentence FLORES+ dev dataset for Southern Uzbek
2. Development of 39,994 parallel sentence pairs across multiple language combinations (uzs-uzn, uzs-en)
3. Fine-tuned NLLB-200 model (lutfiy) optimized for Southern Uzbek translation
4. Post-processing methodology for Arabic script half-space character restoration
5. Open-sourced datasets, models, and evaluation tools

Future work will focus on expanding dataset coverage through additional literary sources and government documents, exploring data augmentation techniques using large language models, and developing more sophisticated orthographic normalization approaches for Arabic script processing.

8 Limitations

Several limitations constrain our current approach. The training dataset size of ~40K sentence pairs, while substantial for a low-resource language, may limit generalization across diverse domains. Our heavy reliance on literary sources potentially biases the model toward formal registers, possibly affecting performance on conversational or technical content. The half-space post-processing solution, while effective, represents a workaround

rather than addressing the underlying tokenizer limitations. Additionally, our evaluation relies primarily on automatic metrics, which may not fully capture translation quality nuances for morphologically complex languages like Southern Uzbek. Human evaluation studies would provide more comprehensive quality assessment.

9 Acknowledgements

We thank the Open Language Data Initiative (OLDI) for supporting this research and David Dale for his valuable guidance throughout the project. The authors thank the Google for Startups Program for providing the computational resources that made this research possible.

References

Afghanistan. 2004. *The Constitution of Afghanistan*. Transitional Islamic State of Afghanistan, Constitutional Commission, Secretariat, Kabul, Afghanistan. Ratified January 26, 2004.

Azizullah Aral. 2024. *Uzbek-English-Turkish-Persian-Pashto Phrasebook*. AkademSpace, Tashkent.

Azizullah Aral. 2025. *Navoi Studies in Afghanistan*. BookmanyPrint, Tashkent.

David Dale. 2022. The first neural machine translation system for the Erzya language. In *Proceedings of the First Workshop on NLP applications to field linguistics*, pages 45–53, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Ehsan Doostmohammadi, Minoo Nassajian, and Adel Rahimi. 2020. Joint Persian word segmentation correction and zero-width non-joiner recognition using BERT. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4612–4618, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ethnologue. 2025a. Southern uzbek. Accessed: 2025-08-10.

Ethnologue. 2025b. Uzbek. Accessed: 2025-08-10.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Google. 2025. Google translate. `https://translate.google.com`. Accessed: August 13, 2025.

Vakul Goyle, Parvathy Krishnaswamy, Kannan Girija Ravikumar, Utsa Chattopadhyay, and Kartikay Goyle. 2023. Neural machine translation for low resource languages.

Fouzia Habibi Aral. 2021. *Brief History of the Uzbek Language*. Ozodiy, Kabul.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Nurullah Oltoy Olim Labib, Azizullah Aral. 2020. Unforgettable - academic seminar on afghan uzbek and turkmen languages. In *Collected Articles*, Kabul. Voja.

Abhinav P M, Ketaki Shetye, and Parameswari Krishnamurthy. 2024. MTNLP-IIITH: Machine translation for low-resource Indic languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 751–755, Miami, Florida, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

U. Tursunov and B. O‘rinboyev. 1982. *History of the Uzbek Literary Language*. O‘qituvchi, Tashkent.

The Kyrgyz Seed Dataset Submission to the WMT25 Open Language Data Initiative Shared Task

Murat Jumashev
jumasheff@gmail.com

Alina Tillabaeva
alinatillabaeva42@gmail.com

Aida Kasieva
aida.kasieva@manas.edu.kg

Turgunbek Omurkanov
omurkanov@gmail.com

Akylai Musaeva
2101.10001@manas.edu.kg

Meerim Emil kyzy
kumaodoru43@gmail.com

Gulaiym Chagataeva
chagataevag.it@gmail.com

Jonathan Washington
jonathan.washington@swarthmore.edu

Abstract

We present a Kyrgyz language seed dataset as part of our contribution to the WMT25 Open Language Data Initiative (OLDI) shared task. This paper details the process of collecting and curating English–Kyrgyz translations, highlighting the main challenges encountered in translating into a morphologically rich, low-resource language. We demonstrate the quality of the dataset through fine-tuning experiments, showing consistent improvements in machine translation performance across multiple models. Comparisons with bilingual and MNMT Kyrgyz-English baselines reveal that, for some models, our dataset enables performance surpassing pretrained baselines in both English–Kyrgyz and Kyrgyz–English translation directions. These results validate the dataset's utility and suggest that it can serve as a valuable resource for the Kyrgyz MT community and other related low-resource languages.

1 Introduction

While machine translation has advanced significantly, progress remains uneven, with morphologically rich, low-resource languages facing substantial obstacles (Goyal et al., 2022). This disparity is particularly acute for the Turkic language family, where agglutinative structures pose unique challenges to standard MT architectures. Kyrgyz (кыргыз тили), a Turkic language with approximately 5.5 million speakers, exemplifies this situation. Despite a growing number of large, automatically-mined parallel datasets (Team et al., 2022), there is a significant shortage of high-quality, human-curated resources essential for building robust and reliable translation systems.

This paper details our contribution to the Open Language Data Initiative (OLDI) shared task: the creation of a high-quality, human-validated English-Kyrgyz dataset of 6,193 sentence pairs. The source text, drawn from diverse scientific domains on Wikipedia, provides rich terminological and syntactic complexity that inherently challenges MT systems. For a low-resource language like Kyrgyz, this complexity exposed a more fundamental obstacle: the absence of standardized scientific vocabulary. This makes high-quality human translation not just a matter of review, but of active linguistic curation. Our translation process, therefore, required developing novel strategies to handle significant terminological gaps and neologisms. This involved our team of native speakers making crucial terminological choices. For instance, we prioritized the native Kyrgyz word "дене" for scientific compounds like "celestial body" ("космостук дене") and "antibody" ("антидене") over the somewhat established Russian calque "тело", a decision that reflects modern usage and enhances both accuracy and naturalness.

Our main contributions are as follows:

1. We contribute the first high-quality, human-validated English-Kyrgyz seed dataset to the OLDI initiative.

2. We provide a detailed analysis of the key linguistic challenges in English-to-Kyrgyz translation, particularly regarding terminological adaptation, and document the effective strategies our team employed.

3. We conduct fine-tuning experiments with four major NMT models (mT5, mBART, M2M100, and NLLB-200) to empirically demonstrate that our high-quality dataset provides consistent performance gains.

Our results confirm that even a modest amount of high-quality parallel data is critical for advancing

Proceedings of the Tenth Conference on Machine Translation, pages 1088–1102
November 8-9, 2025

MT performance for structurally divergent language pairs, particularly when involving a low-resource language.[1]

2 Related Work

Despite Kyrgyz being classified as a low-resource language, recent years have witnessed notable progress in the development of machine translation (MT) systems for this language (Alekseev and Turatali, 2024). The very first machine translation systems for Kyrgyz were rule-based, representing foundational efforts in the field. A key contribution in this area is the open-source finite-state morphological transducer for Kyrgyz developed by Washington et al. (2012). This transducer, developed within the Apertium platform, was a critical component for a prototype Turkish-Kyrgyz machine translation system and laid the groundwork for further language pairs, including an in-progress Kazakh-Kyrgyz system. Alkım and Çebi (2019) proposed a rule-based approach for multilingual translation among four Turkic languages, including Kyrgyz.

Another line of work focusing on Turkic languages is based on neural machine translation (NMT). Mirzakhalov et al. (2021a) trained bilingual models for 22 Turkic languages, including Kyrgyz. In addition to developing baseline systems, this study also introduced a large-scale parallel dataset containing translation pairs for these languages, thereby providing valuable resources for advancing research in Turkic language MT.

Later, the authors released a multi-way multilingual model neural MT model (MNMT) for these languages, showing that the multilingual model outperforms almost all bilingual baselines (Mirzakhalov et al., 2021b).

Fine-tuning multilingual models for low-resource languages has shown considerable promise (Maillard et al., 2023). Notably, Kyrgyz has been incorporated into several multilingual NMT models, such as mT5 (Xue et al., 2021), leading to improved translation quality. A particularly significant contribution to the field is the multilingual NLLB-200 model, trained on 200 languages including Kyrgyz (Team et al., 2022). The primary objective of this work was to provide extensive coverage of low-resource languages within a unified framework.

Another line of research aimed at improving machine translation quality focuses on enhancing tokenization methods. For example, the study by Tukeyev et al. (2020) proposes a tokenization approach based on the Complete Set of Endings (CSE), which reduces vocabulary size and, as demonstrated on Kazakh–English translation tasks, yields better generation quality compared to Byte Pair Encoding (BPE) segmentation. Similarly, Abduali et al. (2025) investigate the Kyrgyz–Kazakh language pair and develop a morphological tokenizers based on the relational segmentation model. The scientific contribution of this article is the creation of the morphological tokenizer for Kyrgyz and Kazakh, as well as fine-tuning experiments with this dataset of the neural model T5-small.

3 Kyrgyz-English Parallel Datasets

Another important line of work in improving machine translation involves the creation of multilingual parallel datasets. Below is a brief overview of open parallel datasets that include the Kyrgyz language.

As part of the TurkLang-7 project (Khusainov and Minsafina, 2021), a corpus of parallel sentences was compiled for Russian and seven Turkic languages. For Kyrgyz, the collection includes 426,190 parallel sentence pairs; however, the final version of the dataset has not been made publicly available.

The NLLB v1 dataset[2] is a large English-Kyrgyz parallel corpus containing 21,360,637 sentence pairs, produced during the development of the NLLB-200 model (Team et al., 2022), where Kyrgyz was included among the languages for which bilingual translation pairs were automatically collected. However, the quality of this dataset is limited, as demonstrated by our manual analysis of the public OPUS sample, which comprises 49 examples[3]. The fully annotated sample that forms the basis for this analysis, including examples of misalignment and various error types, is provided in Appendix A. Our analysis revealed that only 59.18% of the sentences labeled as Kyrgyz were actually written in Kyrgyz. Furthermore, of this reduced Kyrgyz subset, a mere 55.17% were deemed to be accurate and fluent translations. This indicates that only about 32.65% of the original sample represents a high-quality, usable translation pair.

Kyrgyz language was also incorporated into

[1] https://github.com/kyrgyz-nlp/oldi-dataset-experiments

[2] https://opus.nlpl.eu/NLLB/en&ky/v1/NLLB

[3] https://opus.nlpl.eu/sample/en&ky/NLLB&v1/sample

a set of 14 low-resource languages for which the GoURMET project team compiled parallel datasets (van der Kreeft et al., 2022). In total, 14,498 Kyrgyz-English parallel sentences and 23,017 Kyrgyz-Russian parallel sentences were assembled. These sentences were obtained through machine translation followed by editorial validation and control.

The Open Language Data Initiative (OLDI) is a collaborative project that enables language communities, researchers, and developers to contribute to foundational datasets essential for machine learning and natural language processing. At present, OLDI[4] maintains two key datasets: OLDI-Seed and FLORES+.

The OLDI-Seed dataset contains 6,193 English sentences paired with translations into approximately 40 low-resource languages. The English source sentences were drawn from a diverse range of Wikipedia articles covering fields such as biology, astronomy, the arts, history, mathematics, etc. We use the English corpus as a source dataset for Kyrgyz translations.

FLORES+ is an evaluation benchmark consisting of two subsets — a test set of 1012 sentences and a validation set of 997 sentences — each professionally translated into 200 languages by expert linguists.

The X-WMT benchmark (Mirzakhalov et al., 2021b) is a test set designed to evaluate machine translation quality for Turkic languages. It is based on the professionally translated English–Russian corpus from the WMT 2020 News Translation Task. The original news sentences were subsequently translated into eight Turkic languages. For the English–Kyrgyz language pair, the benchmark comprises 500 sentences.

4 Language Description

Kyrgyz (also known as Kirgiz or Kirghiz) is a Turkic language of the Kipchak and/or the South Siberian branch (ISO 639-2: kir, Glottocode: kirg1245). It is spoken primarily in Kyrgyzstan, where it holds official status due to it being the national language. However, it does spread further to Central Asian countries, specifically in Gorno-Badakhshan Autonomous Region of Tajikistan and it is also considered to be a minority language in the Kizilsu Kyrgyz Autonomous Prefecture in Xinjiang, China. And another regional variant of the Kyrgyz language, referred to as Pamiri Kyrgyz, is spoken in northeastern Afghanistan and northern Pakistan.

From a phonetic and phonological standpoint, the Turkic languages most closely resembling Kyrgyz are the southern dialects of Altay, although Kyrgyz also shares significant similarities with Kazakh (Washington et al., 2012). Additionally, the southern varieties of Kyrgyz exhibit distinctive features that align with Uzbek, which are not present in any other Kyrgyz dialects. Approximately, 5.5 million people primarily in Kyrgyzstan consider Kyrgyz their native language.

5 Dataset Translation

5.1 Translation Workflow

Our translation team consisted of six native Kyrgyz speakers, of which two were experienced English to Kyrgyz translators and two were students from the Kyrgyz-English Language Program at Kyrgyz-Turkish Manas University. We followed the OLDI contribution guidelines (Open Language Data Initiative, 2025). The translation workflow was structured into four sequential stages: (1) machine translation using Aitil[5], a Gemma3-based MT service fine-tuned by Ulut Soft LLC (2) manual correction by individual translators, (3) team terminology unification, and (4) consistency review.

At the initial stage, we used the Aitil translation service, which was kindly provided by Ulan Bayaliev and Ulut Soft LLC. At the second stage, translators worked individually to post-edit their assigned batches of machine-translated sentences. This human review process covered all 6,193 sentences, with 99.16% of them being modified. The post-edits addressed a range of common machine translation issues; for instance, our work frequently involved correcting literal translations, improving lexical and terminological choices, and resolving stylistic inconsistencies. To illustrate the nature of these corrections, Table 1 presents several examples.

Following the individual post-editing, we implemented a targeted consistency review process. Our main priority was to ensure high-quality terminological consistency across the entire dataset. To achieve this, translators flagged domain-specific or ambiguous terms for group discussion. Once the team reached a consensus on the translation, the decision was implemented systematically: the

[4]https://oldi.org/

[5]https://translate.mamtil.gov.kg/

Table 1: Examples of machine translations and human post-edits

ID	Sentence
5715	**English:** ...if anyone does something that truly is bad, *it must be unwillingly* or out of ignorance; consequently, all virtue is knowledge. **MT by Aitil:** ...эгер кимдир бирөө чындап эле жаман нерсе жасаса, анда ал муну билбестиктен же *аргасыздан жасашы керек*; демек, бардык жакшылык – бул билим. **Human Post-Edit:** ...эгер кимдир бирөө чындап эле жаман нерсе жасаса, анда ал муну *аргасыздан же билбестиктен улам жасайт*; демек, ар кандай жакшылык – бул билим. **Comment: Syntactic/Stylistic Error:** The MT's output "аргасыздан жасашы керек" is a literal and awkward translation of the English modal structure "it must be". The word order was also unnatural and was corrected in the human edit.
5721	**English:** Although rule by a wise man would be preferable to *rule by law*, the wise cannot help but be judged by the *unwise*... **MT by Aitil:** Акылман адамдын башкаруусу *мыйзамдын башкаруусунан* артык болсо да, акылмандарды *акылсыздар* соттойт... **Human Post-Edit:** Акылмандын башкаруусу *мыйзам үстөмдүгүнөн* артык болсо да, акылмандар сөзсүз *наадандардын* сынына кабылгандыктан... **Comment: Lexical/Terminological Error:** "мыйзам үстөмдүгү" is the correct term for "rule by law". The MT's choice of "акылсыздар" (dumb) for "unwise" was inaccurate; "наадандар" (ignorant/unwise) is more appropriate.
5733	**English:** While the objective of the Pyrrhonists was the attainment of ataraxia, after Arcesilaus the Academic skeptics did not hold up ataraxia as the *central objective*. **MT by Aitil:** Пирончулардын максаты атараксияга жетүү болсо да, Арцелайдан кийин Академиялык скептиктер атараксияны *борбордук максат* катары карманган эмес. **Human Post-Edit:** Пиррончулардын максаты атараксияга жетүү болсо, Аркесилайдан кийин академиялык скептиктер атараксияны *негизги максат* катары көрсөтүшкөн эмес. **Comment: Literal Translation:** "борбордук максат" is a literal translation of "central objective". The more natural term is "негизги максат" (main/primary objective). An extra word ("да") was also removed.

translator who had flagged the term would then perform a search across the entire dataset to update all its occurrences with the agreed-upon translation, ensuring uniformity.

This approach ensured consistency for key concepts. However, we acknowledge two limitations in our overall methodology. First, a comprehensive, sentence-by-sentence peer review was not conducted due to time constraints. Second, the translation workload was unevenly distributed, with one translator contributing the majority of the post-edits:

- Murat: 4910 sentences (79.28%)
- Gulaiym: 614 sentences (9.91%)
- Meerim: 284 sentences (4.58%)
- Elza: 242 sentences (3.91%)
- Akylai: 94 sentences (1.52%)
- Begayim: 49 sentences (0.79%)

5.2 Addressing Problematic Translations

During translation process we turned to dictionaries, both general and technical ones, specific for the given areas, such as, Yudakhin's Russian to Kyrgyz, Kyrgyz to Russian dictionaries (Yudakhin, 1957), (Yudakhin, 1965), Abdiev's English to Kyrgyz, Kyrgyz to English dictionary (Abdiev and Sydykova, 2015) and other dictionaries that are available online at el-sozduk.kg[6]. We also used a collection of Kazakh to Russian and Russian to Kazakh dictionary available at sozdik.kz to verify that international terms (for example, "antibody" which was searched for in Russian: "антитело").

Kyrgyz has certain features that make translation between Kyrgyz and English a challenging task. One of them is the word order. Unlike English, which uses SVO order, Kyrgyz is a SOV type lan-

[6] https://el-sozduk.kg

Table 2: Example of sentence fragmentation

ID	Sentence
4140	**English:** Perhaps the foremost mathematician of the 19th century was the German mathematician Carl Friedrich Gauss, who made numerous contributions to fields such as algebra, analysis, differential geometry, matrix theory, number theory, and statistics.
	Kyrgyz (less effective): Кыязы, 19-кылымдын эң залкар математиги немис математиги Карл Фридрих Гаусс болгон, жана ал алгебра, анализ, дифференциалдык геометрия, матрицалар теориясы, сандар теориясы жана статистика сыяктуу тармактарга көптөгөн салым кошкон.
	Kyrgyz (final): 19-кылымдын эң залкар математиги, кыязы, немис окумуштуусу Карл Фридрих Гаусс болгон. Ал алгебра, анализ, дифференциалдык геометрия, матрицалар теориясы, сандар теориясы жана статистика сыяктуу тармактарга зор салым кошкон.

guage. Although the placement of the verb does not pose any issues, it becomes more difficult the more complex the sentence is. Syntactic functions (e.g., adverbial phrases) also have to be displaced, which makes translation more complicated, as it becomes harder to keep the sentence eligible for the readers without losing the natural flow of the language. (Sankaravelayuthan, 2019). Our translators have also encountered this particular problem, that is dealt with sentence fragmentation. Another feature is that Kyrgyz is agglutinative. It forms words through the sequential addition of morphemes while preserving their original spelling and pronunciation. This linguistic structure allows for the creation of an extensive range of word forms, while English, being an analytical language, relies heavily on such features as auxiliary words, modal verbs and dependent clauses rather than inflection to express grammatical relationships (Kara, 2003).

While the challenges mentioned before were expected, our translators have encountered another two major problems: long, compound-complex sentences and terminological gaps and neologisms. Those problems significantly hindered the translation process. The first problem was dealt with sentence fragmentation, a syntactic transformation technique (Shermatova, 2010). This technique involves restructuring a single complex sentence from the source language into two or more sentences in the target language. The goal is to maintain a natural linguistic flow, and to preserve the original's clarity, emphasis, and stylistic integrity. An example of such fragmentation is listed in Table 2.

The less effective version contains two coordinated clauses with the conjunction "жана" ("and"), presents two critical problems: stylistical redundancy (the repetition of the word "математиги") and syntactic overload (while technically a single grammatical unit, the sentence is long and burdensome, combining two distinct ideas with a simple conjunction, which weakens its clarity and impact). This final resolves both issues. The redundant term is eliminated by using a more appropriate synonym ("окумуштуусу" - "scientist") in the first sentence and allowing the subject to be implied in the second. Most importantly, the structure is clear, emphatic, and stylistically natural in Kyrgyz. The main challenges were translational gaps and neologisms, especially in scientific and technical fields such as genetic engineering and astronomy. Lacking equivalents, we often used transliteration and calquing, typically via Russian—a legacy of the Soviet era when Russian dominated science and education. As many specialists are bilingual, adopting Russian-based forms speeds comprehension (Table 3).

The terms "knockout" and "extinction" are highly specific neologisms. With "knockout" creating a descriptive Kyrgyz phrase (e.g., "генди өчүрүү"–"gene deactivation") would be less precise and more cumbersome. Coining an entirely new Kyrgyz term would likely be unintelligible to specialists who are already familiar with the concept through international, often Russian-language, literature. Therefore, the most effective strategy is the transliteration of the international term (or calquing of Russian term) as "нокаут". This direct adoption is efficient and aligns with the established scientific vocabulary (Zaid et al., 2008).

For accuracy and consistency, we adopted the international term "экстинкция", sourced from an English–Russian astronomical dictionary (Murtazov, 2010), as it aligns with existing literature and is more recognizable to Kyrgyz readers. The problems with scientific terms would not have posed

Table 3: Examples of terminology mapping

ID	Sentence
3986	**English:** In a simple **knockout** a copy of the desired gene has been altered to make it non-functional **Kyrgyz:** Жөнөкөй **нокаутта** керек болгон гендин көчүрмөсү иштебей калышы үчүн өзгөртүлөт.
3506	**English:** In astronomy, **extinction** is the absorption and scattering of electromagnetic radiation by dust and gas between an emitting astronomical object and the observer **Kyrgyz:** Астрономияда **экстинкция** – бул нур чыгаруучу астрономиялык объект менен байкоочунун ортосундагы чаң жана газ тарабынан электромагниттик нурлануунун жутулушу жана чачырашы.

such substantial problems if not for the major linguistic flaw that calquing had created. In most of the cases calquing was justified, as it provided a more specific and in-depth understanding of the phenomenon, in some it resulted in a borrowed word replacing a perfectly suitable native synonym. The primary example is showcased in Table 4.

The case with "bodies" highlights this protrusion. The official dictionary (Eshbaev and Eshbaeva, 2023) suggests using the Russian calque "тело," whereas we observe that this term is now often replaced by the Kyrgyz word "дене." This modern usage is evident in real-world examples from publications such as BBC News Kyrgyz[7], which uses "антидене" for "antibody," and Nazar News[8], which uses "космостук дене" for "cosmic body." During translation, our team gave priority to the latter option, as it better reflects the language usage.

6 Translation Experiment

6.1 Model Training

To demonstrate the quality of the collected dataset, we fine-tuned several seq2seq machine translation models on the gathered 6,193 parallel sentences and evaluated their performance on the FLORES+ benchmark (Team et al., 2022; The Open Language Data Initiative, 2024). We also evaluate our model on the Turkic X-WMT benchmark to compare its performance with the bilingual and multilingual baselines proposed for Kyrgyz–English machine translation.

For fine-tuning, we utilized the Transformers library (Wolf et al., 2020). Across all experiments, we employed the AdamW optimizer with a learning rate of 0.0001 and trained for 6 epochs. The training was conducted on a single T4 GPU with 16 GB of memory.

For each model, we take the same base instance and fine-tune it twice independently — once for translating from English to Kyrgyz (en → ky) and once for translating from Kyrgyz to English (ky → en). The models selected for fine-tuning included:

mT5-base is a multilingual transformer model based on the T5 architecture. It is pretrained on a denoising objective across 101 languages, including Kyrgyz. While mT5 supports Kyrgyz at the pretraining stage, it requires fine-tuning on translation tasks to perform effective machine translation into and from Kyrgyz. (Xue et al., 2021)

mBART-large is a seq2seq transformer extending the BART model to multilingual settings. However, the base mBART-large model does not include Kyrgyz in its pretrained vocabulary, limiting its direct applicability to Kyrgyz translation without additional fine-tuning and vocabulary extension. (Tang et al., 2020)

M2M100 is a multilingual seq2seq model developed by Facebook supporting direct translation between 100 languages without relying on English as a pivot. Notably, Kyrgyz is not included in the 100 languages covered by M2M100, thus the model cannot translate Kyrgyz "out of the box." (Tang et al., 2020)

NLLB (No Language Left Behind) is a large-scale multilingual seq2seq model developed by Meta, designed to improve translation quality, particularly for low-resource languages. It supports over 200 languages, including Kyrgyz. NLLB can translate to and from Kyrgyz with high quality without requiring additional fine-tuning, making it well-suited for applications involving Kyrgyz language translation. (Team et al., 2022)

[7] https://www.bbc.com/kyrgyz/articles/c0lye5ngjwxo

[8] https://nazarnews.org/posts/zherdin-kagyilyishuusu-eki-asman-tiregen-asteroid-zherge-zhakyindadyi

Table 4: An example of an unnecessary calque

ID	Sentence
3550	**English:** Once it became clear that Earth was merely one planet amongst countless **bodies** in the universe, the theory of extraterrestrial life started to become a topic in the scientific community. **Kyrgyz:** Жер – ааламдагы сансыз асман **денелеринин** арасындагы катардагы эле бир планета экени айкын болгондон кийин, жерден тышкаркы жашоо теориясы илимий коомчулукта талкуулана баштаган.

6.2 Vocabulary expansion

For the models that were not pretrained on Kyrgyz (mBART and M2M100), we expanded the token vocabulary by training a SentencePiece model (Kudo and Richardson, 2018) on the Kloop corpus, a dataset of Kyrgyz news articles (kyrgyz-nlp, 2024). This allowed us to extend the vocabulary by 14,466 byte-pair encoding (BPE) tokens (Sennrich et al., 2016).

6.3 Metrics

To evaluate generation quality, we employ two variations of the ChrF metric: ChrF (Popović, 2015) and ChrF++ (Popović, 2017), both of which are well suited for morphologically rich languages. While prior work has shown that ChrF++ correlates more strongly with human judgments, we also report ChrF scores to enable direct comparison with the Bilingual and MNMT baselines (Mirzakhalov et al., 2021b). For this purpose, we use the implementation provided in the SacreBLEU toolkit[9], adopting the same parameter configuration as in the MWMT study for the computation of the ChrF metric to ensure consistency and comparability of results.

6.4 Experiment Results

The tables present the training results of our models, evaluated on two benchmarks: FLORES+ and X-WMT.

We compare the outputs of our fine-tuned models with its pretrained versions (**pretrained baselines**) on the FLORES+ benchmark to evaluate the impact of our dataset on translation quality. The results are reported separately for the two translation directions: English→Kyrgyz (Table 5) and Kyrgyz→English (Table 6). The results of the X-WMT benchmark are summarized in Table 7.

Fine-tuning on our dataset substantially improves generation quality for all models in both translation directions. (Table 5 and Table 6)

NLLB-200 achieved the highest performance in all the evaluated models. Although the models were already pre-trained on a large Kyrgyz corpus, the addition of even a relatively small dataset like ours can yield a noticeable improvement in translation performance.

When comparing translation directions, Kyrgyz–English translations (Table 6) outperforms English–Kyrgyz translations (Table 5) for all models except M2M100. This asymmetry reflects the stronger representation of English in the pretraining corpora, which enables more reliable decoding into English. The exception of M2M100 can be attributed to its non-English-centric pretraining design.

Manual validation of system outputs, however, revealed that automatic evaluation scores tend to overestimate real-world usability. Despite relatively high ChrF and ChrF++ scores, the translations still contained a considerable number of critical errors. In particular, M2M100 outputs suffered from frequent tokenization and word-concatenation issues, likely caused by suboptimal adaptation of the tokenizer. More generally, all systems struggled with morphological accuracy, especially the generation of correct suffixes. This challenge stems from the agglutinative nature of Kyrgyz, where the high variability of word endings makes exact surface realization difficult.

The performance of pretrained baselines largely depends on whether Kyrgyz was included in their pretraining data and the model's intended use. For example, although mT5's pretraining corpus contains Kyrgyz, it cannot perform machine translation "out of the box" and requires fine-tuning. Neither mBART nor M2M100 was pretrained on Kyrgyz; nevertheless, M2M100 achieves comparatively stronger baseline results, as we extended the model to Kyrgyz by introducing a new language token and initializing its embeddings from Kazakh, a closely related language. Finally, NLLB-200, which was trained on a large Kyrgyz corpus, achieves high

[9] https://github.com/mjpost/sacrebleu

Table 5: Performance comparison on the FLORES+ English-Kyrgyz translation benchmark.

	mT5	mBART	M2M100	NLLB-200
fine-tuning	ChrF: 30.57 ChrF++: 26.53	ChrF: 34.85 ChrF++: 30.82	ChrF: 40.77 ChrF++: 33.91	ChrF: 49.37 ChrF++: 44.62
pretrained baselines	ChrF: 0.20 ChrF++: 0.63	ChrF: 0.73 ChrF++: 1.76	ChrF: 10.88 ChrF++: 9.10	ChrF: 44.56 ChrF++: 40.21

Table 6: Performance comparison on the FLORES+ Kyrgyz–English translation benchmark.

	mT5	mBART	M2M100	NLLB-200
fine-tuning	ChrF: 36.82 ChrF++: 34.22	ChrF: 36.23 ChrF++: 33.88	ChrF: 38.77 ChrF++: 36.19	ChrF: 51.77 ChrF++: 49.34
pretrained baselines	ChrF: 1.84 ChrF++: 1.81	ChrF: 0.84 ChrF++: 1.22	ChrF: 13.29 ChrF++: 10.58	ChrF: 49.85 ChrF++: 47.48

Table 7: Performance comparison on the X-WMT dataset (ChrF). Values that exceed the performance of the Multilingual MNMT baseline are highlighted in bold. For the ChrF score, we did not reproduce the baselines ourselves, but used the results reported in the paper.

	mT5	mBART	M2M100	NLLB-200	Bilingual XWMT	Multilingual MNMT
en-ky	28.32	30.92	**36.41**	**47.61**	27	34
ky-en	34.69	33.68	36.10	**47.84**	29	39

performance without additional fine-tuning.

All of our trained models outperform the bilingual MWMT baselines (Table 7). Among them, the multilingual MNMT model is surpassed only by NLLB-200 in both translation directions, and by M2M100 in the English→Kyrgyz direction, demonstrating the competitive performance of our models across different translation setups.

7 Conclusion

This study, part of the OLDI initiative, contributed 6,193 English-Kyrgyz sentence pairs and showed that even modest, carefully curated data can improve neural machine translation for low-resource, morphologically complex languages. In this work, we addressed three key challenges: adapting to agglutinative morphology, restructuring complex subordination, and maintaining terminological consistency. Our preference for native Kyrgyz terms over Russian calques reflects contemporary usage and contributes to natural language evolution.

Fine-tuning mBART, M2M100, mT5, and NLLB-200 models on our dataset yielded consistent performance gains, validating the value of linguistically-informed data. However, the manual evaluation of model translations, highlight the need for further research to achieve production-quality NMT for Kyrgyz. Our collaborative workflow—combining MT, manual correction, and consistency review—offers a replicable methodology for other low-resource languages. Future work involves expanding the dataset and integrating community feedback. This research provides concrete resources and a methodological framework for community-driven language technology development, balancing technical advancement with cultural-linguistic authenticity.

8 Acknowledgments

We would like to thank Elza Mambetakunova and Begaiym Mamatova for helping us with the translations. We are also grateful to Ulan Bayaliev and Ulut Soft LLC for providing access to their Aitil translator, which served as a valuable resource for this work.

References

Taalaibek Abdiev and Lira Sydykova. 2015. *Anglische-Kyrgyzcha, Kyrgyzcha-anglische syozduk [English-Kyrgyz, Kyrgyz-English Dictionary]*. Avrasya Press, Bishkek.

Balzhan Abduali, Ualsher Tukeyev, Zhandos Zhumanov, and Nella Israilova. 2025. Study of Kyrgyz-Kazakh Neural Machine Translation. In *Proceedings of the 17th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2025)*, pages 272–283, Kitakyushu, Japan. Springer.

Anton Alekseev and Timur Turatali. 2024. KyrgyzNLP: Challenges, progress, and future. *arXiv preprint*, arXiv:2411.05503.

Emel Alkım and Yalçın Çebi. 2019. Machine translation infrastructure for Turkic languages (MT-Turk). *The International Arab Journal of Information Technology*, 16(3).

A. A. Eshbaev and Ch. A. Eshbaeva. 2023. *Russian-Kyrgyz Explanatory Dictionary of Frequently Occurring Terms in Medicine*. Kyrgyz Encyclopedia and Terminology Center, Bishkek.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

D. S. Kara. 2003. *Kyrgyz*. Languages of the World/Materials. Lincom Europa, Munich.

Aidar Khusainov and Alina Minsafina. 2021. First results of the "TurkLang-7" project: Creating Russian-Turkic parallel corpora and MT systems. In *Proceedings of the International Workshop on Computational Models in Language and Speech (CMLS 2021)*. CEUR-WS.org. Held as part of the 10th International Conference on Analysis of Images, Social Networks and Texts (AIST 2021).

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

kyrgyz-nlp. 2024. Kloop corpus. `https://github.com/kyrgyz-nlp/kloop-corpus`.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Behzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021a. A large-scale study of machine translation in the Turkic languages. *arXiv preprint*, arXiv:2109.04593.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Bekhzodbek Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, Francis Tyers, Orhan Firat, John Licato, and Sriram Chellappan. 2021b. Evaluating multiway multilingual NMT in the Turkic languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 518–530, Online. Association for Computational Linguistics.

A. K. Murtazov. 2010. *Russian-English Astronomical Dictionary*. Ryazan State Pedagogical University, Ryazan.

Open Language Data Initiative. 2025. Contribution guidelines. Accessed: 2025-08-13.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

R. Sankaravelayuthan. 2019. Word order in translation. *Language in India*, 19(4):196–206.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Feruza S. Shermatova. 2010. *Features of the translation of syntactic stylistic devices from English to Kyrgyz*. Ph.D. thesis, Kyrgyz-Russian Slavic University, Bishkek.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel

Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

The Open Language Data Initiative. 2024. Flores+ dataset. https://huggingface.co/datasets/openlanguagedata/flores_plus. Accessed: 2025-09-04.

Ualsher Tukeyev, Aidana Karibayeva, and Zh. Zhumanov. 2020. Morphological segmentation method for Turkic language neural machine translation. *Cogent Engineering*, 7(1):1856500.

Peggy van der Kreeft, Sevi Sariisik, Wilker Aziz, Alexandra Birch, and Felipe Sánchez-Martínez. 2022. GoURMET – machine translation for low-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 225–226, Ghent, Belgium. European Association for Machine Translation.

Jonathan North Washington, Mirlan Ipasov, and Francis M. Tyers. 2012. A finite-state morphological transducer for Kyrgyz. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2244–2248, Istanbul, Turkey. European Language Resources Association (ELRA).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

K. K. Yudakhin, editor. 1957. *Russko-kirgizskij slovar' [Russian-Kyrgyz Dictionary]*. State Publishing House of Foreign and National Dictionaries, Moscow.

K. K. Yudakhin. 1965. *Kirgizsko-russkij slovar' [Kyrgyz-Russian Dictionary]*. Sovetskaja Enciklopedija, Moscow.

A. Zaid, H.G. Hughes, E. Porcheddu, and F. Nicholas. 2008. *Glossary of biotechnology for food and agriculture (Russian Edition)*, volume 9 of *FAO Research and Technical Paper*. Food and Agriculture Organization of the United Nations (FAO), Rome. Translated into Russian by G. Kamarova, T. Gavrilenko, I. Anisimova, O. Antonova, O. Kuznetsova, and S. Kharitonov.

A Qualitative Analysis of the Official OPUS Sample for NLLB v1 (En-Ky)

Table 8: Sentence pairs drawn from the public sample of the NLLB v1 En-Ky corpus provided by the OPUS project. Each pair is manually evaluated for language identification (`Is Kyrgyz`) and translation accuracy (`Is Correct`). The `Comments` column details the types of errors found.

En	Ky	Is Kyrgyz	Is Correct	Comments
And if you call them to guidance, they do not listen.	Эгер аларды түз жолго чакырсаң угушпайт.	Yes	Yes	
(It is only) a provision in this world, then to Us is their return.	Бул дүйнөдө жыргал, андан соң алардын Бизге кайтуулары бар.	Yes	Yes	
Are you [more] satisfied with the life of this world, rather than the Hereafter?	Дүнё уларга, Охират бизга эканига рози эмасмисан?!	No	No	
If GOD wills, He can take away their hearing and their eyesight.	Эгер Алла каалаганда, алардын угуусун да, көрүүсүн да алып коймок эле.	Yes	Yes	
and you will be among those brought near."	Ва, албатта, менинг яқин кишиларимдан бўлурсизлар," деди."	No	No	
and We see it to be near.	Биз эса уни яқин деб билурмиз.	No	No	
But his people's only answer was, "Bring down upon us God's chastisement, if thou art a man of truth!"	Ошондо анын коомунун жообу: "Эгер чынчылдардан болсоң, Алланын азабын бизге келтирчи!" - дегенинен башка болбоду.	Yes	Yes	
That we might follow the magicians if they are the victorious?"[1]	Агар сеҳргарлар ғолиб бўлиб чиқсалар, эҳтимол бизлар ўшаларга эргашурмиз," дейилди.	No	No	
For the wrong-doers there will be no helpers.	золим кимсалар учун бирон ёрдамчи бўлмас!."[20]	No	No	
(And your abode is the Fire, and there is none to help you.)	Жойингиз жаҳаннамдир ва сизларга ҳеч ёрдамчилар йўқ.	No	No	
But worship Allah alone and be among the grateful" [Qur'an, 39:66].	Аллахка гана сыйын жана шүгүр кылуучулардан бол!"	Yes	Yes	

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
Say, 'They are appointed periods of time for (general convenience of) people and for determining the time of Pilgrimage.	Айт: "Булар адамдарга убакытты жана ажылыкты белгилөө куралы."	Yes	Yes	
And remember the name of your Lord, morning and evening.	Жана эртели-кеч Роббиңдин атын зикир кыл!	Yes	Yes	
Indeed, there has come to you a bearer of glad tidings and a warner.	Ошентип, албетте, силерге куш кабар берүүчү жана коркутуучу келди.	Yes	No	Ошентип is extra
Call upon your helpers, other than Allah, to assist you, if you are true.	Жана эгер чынчыл болсоңор, Аллахтан башка күбөлөрүңөрдү (ишенген жардамчыларыңарды) чакыргыла.	Yes	No	has a comment in parentheses
Protect yourself and your family from the fire of hell."	Өзүңөрдү жана үй-бүлөңөрдү Тозоктун отунан коргогула!" - деп буюрулган.	Yes	No	- деп буюрулган. is extra
And that My punishment is a painful retribution.	менинг азобимдан азоблангани...	No	No	
- that Day, man will remember, but how [i.e.,	Адам ал күнү ойлонуп-эстейт, бирок (бул) эстөөдөн ага эмне пайда?	Yes	No	the English part is incomplete
Do they expect anything but the likes of the days of those who passed away before them?	Алар (капырлар) Кыямат Күнүнүн капилет келишин күтүп жатышабы?Анын белгилери келди.	Yes	No	has a comment in parentheses; Анын белгилери келди. is extra
Remember the name of your Lord morning and evening.	Жана эртели-кеч Роббиңдин атын зикир кыл!	Yes	Yes	
2:55 - Who were the sons of the servants of Solomon?	Эзра 2: 55 - Сулаймандын кулдарынын уулдары деген кимдер эле ?	Yes	Yes	
Why did you kill them, if you are telling the truth?"	Силер чынчыл болсоңор анда эмне үчүн аларды өлтүрдүңөр?!"	Yes	Yes	

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
Indeed, Gehenna is your recompense, and the reward of those who follow you, an ample recompense.	Бас улардан ким сенга эргашса, у ҳолда, шак-шубҳасиз, жаҳаннам сизларга етарли жазо бўлур!	No	No	
WE have created them of that which they know.	Албетте, Биз аларды өздөрү билген нерседен жаратканбыз.	Yes	Yes	
in which they will remain forever, and will not find any guardian or helper.	Алар ал жерде (тозокто) дос жана жардамчы таппай, түбөлүккө калышат	Yes	No	has a comment in parentheses
Mankind were not but one nation (community), but they differed (later).	(Аввалда) одамлар фақат бир миллат (яъни, бир динда) бўлган эдилар.	No	No	
The Fire will be your refuge, and you will have no helpers.	Силердин жайыңар - от болот жана силерге эч кандай жардамчылар болбойт.”	Yes	Yes	
(The Day that Allah will not disgrace the Prophet (Muhammad) and those who believe with him.	Бир кундаки, унда Аллоҳ Набийни ва у билан бирга бўлган иймон келтирганларни шарманда қилмас.	No	No	
Thus God makes clear His Revelations to you, that you may be thankful.” (5:89).	Шүкүр кылууңар үчүн Алла силерге Өз аяттарын мына ушундай ачык-айкын баян кылат.	Yes	Yes	
And that My chastisement is the painful chastisement.	менинг азобимдан азоблангани...	No	No	
And what will make you know what Al-Qari’ah is?	(8) ”Сижжийн” қандай нарса эканини сенга нима билдирди?!	No	No	
(and it will be said to them,).This is the Fire you used to deny	52: 15 ”Силер жалганга чыгарган от - мына ушул.	Yes	No	the Kyrgyz part is incomplete
But surely now has come to you a bearer of glad tidings and a Warner.	Ошентип, албетте, силерге куш кабар берүүчү жана коркутуучу келди.	Yes	Yes	

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
call your witnesses, besides Allah, if you are telling the truth.	Жана эгер чынчыл болсоңор, Аллахтан башка күбөлөрүңөрдү (ишенген жардамчыларыңарды) чакыргыла.	Yes	No	has a comment in parentheses
When a warner came to them, however, it only increased their aversion.	Бирок аларга бир эскертүүчү-коркутуучу келгенде, (бул алардын) жек көрүүлөрүнөн башкасын көбөйткөн жок.	Yes	No	has a comment in parentheses; mistranslation due to incorrect pluralization of the abstract noun жек көрүү (hatred).
But we see it (quite) near.”	Биз эса уни яқин деб билурмиз.	No	No	
Cretans and Arabs, we hear them in our own tongues speaking of the mighty deeds of God.”	криттиктер жана аравиялыктар, өз тилибизде алардын Кудайдын улуу иштери жөнүндө айтып жаткандарын угуп жатабыз.”	Yes	Yes	
And mention the name of your lord (in prayer) morning and evening (Al-Quran 48:9)	Жана эртели-кеч Роббиңдин атын зикир кыл!	Yes	Yes	
Rather, every one of them wishes that he should be given unrolled (divine) scriptures	Балки, улардан ҳар бир киши ўзига очилган саҳифалар берилишини ирода қилар.	No	No	
did you kill them, if you are true in your claim?”	Агар ростгуй булсаларингиз, нимага уларни улдирдингиз?” деб айт.	No	No	
Say: ”You can only expect for us one of two good things; while we wait for you to be afflicted by God with a retribution from Him, or by our hands.	Биз силерге Алладан жаза келерин же биз аркылуу силерге бир азап жиберерин күтүп жатабыз.Анда, силер күткүлө, силер менен чогуу биз да күтөбүз” - деп айт.	Yes	No	completely misaligned fragment
We are only waiting for the order of Allah the Almighty.	Биз анын акысын бир гана Аллаху тааладан күтөбүз,” - деди.	Yes	No	completely misaligned fragment

Table 8: (Continued)

En	Ky	Is Kyrgyz	Is Correct	Comments
Your abode is the fire, and you shall not have any helpers.”	Жойингиз жаҳаннамдир ва сизларга ҳеч ёрдамчилар йўқ.	No	No	
But Iblis (did it not), he refused to be with those who made obeisance.	Ичине Жин кирген адамдарга окшоп калышыптырго.	Yes	No	completely misaligned fragment
hath prepared for them a goodly recompense.	учун улуғ мукофот (яъни, жаннат) тайёрлаб қўйгандир.	No	No	
Then bring your book, if you are truthful.” (37:149-157).	Uzbek / Ozbekcha / Özbekçe Агар ростгўйлардан бўлсангиз, китобингизни келтиринг!	No	No	
When this was clearly shown to him he said: ”I know now that God is able to do all things.”	Ал ага (булар) ап-ачык белгилүү болгондон кийин (мындай) деди: ”Аллахтын бүт нерсеге күч жеткирүүчү экенин билем.”	Yes	No	has a comment in parentheses; awkward literal translation: The phrase ”able to do all things” (implying omnipotence) is rendered as күч жеткирүүчү, which literally means ”the one who delivers/supplies power.” A more idiomatic and accurate translation would be Аллахтын бүт нерсеге күчү жетет экенин билем.
For the wrong-doers there will be no helpers.	Энди золим кимсалар учун бирон ёрдамчи бўлмас!	No	No	
Those people, it has not been for them to enter them except fearing.	бу кимсалар учун ундай жойларга фақат қўрққан ҳолларида кириш жоиз эди-ку.	No	No	

SMOL: Professionally translated parallel data for 115 under-represented languages

Isaac Caswell[ʕ*] **Elizabeth Nielsen**[ʕ*] **Jiaming Luo**[ʕ] **Colin Cherry**[ʕ]
Geza Kovacs[ʕ] **Hadar Shemtov**[ʕ] **Partha Talukdar**[ʕ] **Dinesh Tewari**[ʕ]
Baba Mamadi Diane[ɔ]$_{nqo}$ **Djibrila Diane**[ɔ]$_{nqo}$ **Solo Farabado Cissé**[ɔ]$_{nqo}$
Koulako Moussa Doumbouya[ð]$_{nqo}$ **Edoardo Ferrante**[ʒ]$_{lij}$ **Alessandro Guasoni** [ʒ]$_{lij}$
Christopher Homan[π]$_{dje}$ **Mamadou K. Keita**[π]$_{dje;mos}$ **Sudhamoy DebBarma**[α]$_{trp}$ **Ali Kuzhuget**[θ]$_{tyv;ru}$
David Anugraha[ð]$_{id}$ **Muhammad Ravi Shulthan Habibi**[æ]$_{id}$ **Sina Ahmadi** [ü]$_{ku++}$
Anthony Munthali[ŵ]$_{tum}$ **Jonathan Mingfei Lau**[ʕ]$_{粵}$ **Jonathan Eng**[ʕ]$_{粵}$
[ʕ]Google {Research, Deepmind} [ð]Stanford University [ɔ]NKo USA INC
[ʒ]Conseggio pe-o patrimònio linguistico ligure [æ]Universitas Indonesia [ü]University of Zurich
[π]Rochester Institute of Technology [θ]tyvan.ru

$_{nqo;lij;dje;mos;trp;tyv;id;ku++;tum;粵}$ Led volunteer contributions for NKo, Ligurian, Zarma, Mooré, Kokborok, Tuvan, Russian, Indonesian, Kurdish languages, Tumbuka, and Cantonese, respectively

[*]{icaswell, eknielsen}@google.com

Abstract

We open-source SMOL (*Set of Maximal Overall Leverage*),[1] a suite of training data to unlock machine translation for low-resource languages. SMOL has been translated into 123 under-resourced languages (125 language pairs),[2] including many for which there exist no previous public resources, for a total of 6.1M translated tokens. SMOL comprises two sub-datasets, each carefully chosen for maximum impact given its size: SMOLSENT, a set of sentences chosen for broad unique token coverage, and SMOLDOC, a document-level resource focusing on a broad topic coverage. They join the already released GATITOS for a trifecta of paragraph, sentence, and token-level content. We demonstrate that using SMOL to prompt or fine-tune Large Language Models yields robust CHRF improvements. In addition to translation, we provide factuality ratings and rationales for all documents in SMOLDOC, yielding the first factuality datasets for most of these languages.

1 Introduction

There exist no professionally-translated data for most of the world's 7000 or so languages, rendering tasks like Machine Translation near impossible. High-quality data is needed. However, it is not clear how best to use a limited budget for an expensive task like professional translation. As shown by the GATITOS dataset (Jones et al., 2023), word-level translations provide large benefits to translation quality for low-resource languages at the lowest cost. However, gains quickly saturate, as single tokens are not very expressive. Sentence-level data is better for a model once token-level data saturates, but it has much more inherent redundancy; and document-level data is even more effective...and more redundant.

In this work, we release the SMOL dataset, which provides professionally translated sentence- and document-level data for 123 LRLs (125 language pairs). SMOL contains two sub-datasets:

- **SMOLSENT**: 863 English sentences covering 5.5k of the most common English tokens,[3] professionally translated into 90 languages.
- **SMOLDOC** 584 English documents covering a wide range of topics, domains, and tokens, generated by an LLM and professionally translated into 103 languages.

We demonstrate the utility of these data for fine-tuning and prompting LLMs for translation, and provide factuality annotations for all documents.

[1] 🤗 **google/smol**

[2] Experiments are mostly on a subset of 115 languages, before volunteer translations of additional languages finished. The paper title reflects this.

[3] In this paper, 'token' refers to typographic units as an approximation to words, not subword tokens from a model's vocabulary.

Proceedings of the Tenth Conference on Machine Translation, pages 1103–1123
November 8-9, 2025

2 Related work

There are not many training datasets with human-translated data for Low-Resource Languages (LRLs), where we operationally define LRL as any language beyond the first 100 supported by most traditional crawls and MT providers (enumerated in Appendix section A).

Tatoeba (Tiedemann, 2020) is probably the most multilingual, but it is made of volunteer contributions and of unclear quality. The GATITOS dataset (Jones et al., 2023) consists of a 4000-entry lexicon translated into 170 LRLs, but is only token-level. Most similar to the present work, NLLB-SEED is a high-quality, sentence-level training set of 6k sentences selected from English Wikipedia and professionally translated into 44 LRLs (Team et al., 2022). There are also several professionally-translated evaluation sets, namely FLORES-101 and FLORES-200 (Goyal et al., 2022; Team et al., 2022), and NTREX (Federmann et al., 2022a).

While highly multilingual, professionally translated training data is rare, there is a growing number of bottom-up community data sources organized through research collectives like Masakhane (∀ et al., 2020), Turkish Interlingua (Mirzakhalov et al., 2021a,b), and GhanaNLP (Azunre et al., 2021a); and conferences and workshops like AfricaNLP, AmericasNLP (Mager et al., 2021) and ArabNLP. These datasets are usually generated by researchers fluent in the languages, and are therefore especially high quality. In addition to providing datasets, such efforts frequently also provide models and baselines, or even public interfaces, like the Khaya Translator Web App[4] by GhanaNLP for West African languages, and the lesan.ai[5] translation website for Ethiopian languages.

Participation is especially strong from the African continent, including corpora and models for pan-East-African languages (Babirye et al., 2022), languages from the Horn of Africa (Hadgu et al., 2022), Ethiopian languages (Teferra Abate et al., 2018; Gezmu et al., 2021), Ugandan languages (Akera et al., 2022), Emakhuwa (Ali et al., 2021), South-African languages (Eiselen and Puttkammer, 2014), Setswana and Sepedi (Marivate et al., 2020), Yorùbá (Adelani et al., 2021b,a), Oshiwambo (Nekoto et al., 2022), Igbo (Ezeani et al., 2020), Zulu (Mabuya et al., 2021), Twi (Azunre et al., 2021b), Gbe (Hacheme, 2021), Bambara (Tapo et al., 2021), and Fon (Emezue and Dossou, 2020). Outside of Africa, corpora have been created for languages of the Americas, including for four indigenous languages of Peru in Bustamante et al. (2020), the numerous results on the largely South- and Central American languages from the first AmericasNLP conference (Mager et al., 2021), and the Inuktitut language of Canada (Joanis et al., 2020). Datasets for lower-resourced languages of India have also sprung up, including the 13-language PMIndia (Haddow and Kirefu, 2020), and datasets focused on languages of the Northeast like Mizo (Thihlum et al., 2020), Khasi (Laskar et al., 2021) and Assamese (Laskar et al., 2020). Further West, PARME (Ahmadi et al., 2025) has provided some of the first human-translated content for Kurdish and Iranian languages. Finally, a variety of such datasets and models are available for public use on HuggingFace[6] or Zenodo.[7]

In addition to professionally translated data, there are also several web-crawled datasets for LRLs, including MADLAD (Kudugunta et al., 2023), OSCAR (Ortiz Suárez et al., 2019), Glot500-C (Imani et al., 2023), NLLB (Team et al., 2022), and the Bloom library (Leong et al., 2022).

3 Text Selection

Translation requires significant investment and can't be easily re-done, so great care needs be put into carefully choosing sentences to translate. For both sub-datasets SMOLDOC and SMOLSENT, selection or generation of source text is done in English. Selecting only English has clear biases, but also has advantages—most notably, for N languages, it requires N times less work to quality control. Future work should consider focusing on non-English sources.

3.1 SMOLSENT: Token Set Cover

Our basic motivation for creating SMOLSENT was to help models overcome vocabulary issues, which are common for the lowest-resource languages (Nielsen et al., 2025; Bapna et al., 2022). Therefore, we frame this as a set-cover problem, and pick the smallest set of sentences (from CommonCrawl[8]) that covers the largest set of target tokens. The tokens we chose to cover (the *target set*) were

[4] https://ghananlp.org/project/translator-webapp/

[5] https://lesan.ai/translate

[6] https://huggingface.co/datasets?multilinguality=multilinguality:translation&task_categories=task_categories:translation

[7] https://zenodo.org/communities/africanlp/

[8] https://commoncrawl.org/ we use all available snapshots as of August 20, 2022

Method	ChrF
Random	30.5
Token set-cover	**31.7**
N-gram DWD	30.0
Embedding DWD	27.5

Table 1: Held-out CHRF for data selection approaches

the English side of GATITOS, as well as the most common 2,500 tokens from an English web crawl. Set cover is NP-hard, so we approximate it with a greedy algorithm that iteratively picks the sentence with the highest *coverage percent*, defined as the percentage of its tokens that are in the target set.

Preliminary work on Token Set-Cover To evaluate the token set-cover approach, we started by selecting data from existing web-scraped parallel data. We pretrain a multilingual Neural Machine Translation (NMT) model on parallel data from 294 language pairs from MADLAD-400 (Kudugunta et al., 2023), with nine languages held out to simulate LRLs. We fine-tune this model on sets of existing parallel data in each of the held-out languages, and evaluate on FLORES-200. Details on the experimental set-up in Appendix B.1.

In addition to Greedy Token Set-Cover, we explore two methods that balance data diversity and data quality. First, we implement Ambati et al. (2011)'s 'density-weighted diversity' (DWD) metric, which is an n-gram based metric for diversity and quality. Second, we implement an embedding-based version of DWD, which takes the weighted harmonic mean of perplexity under the Palm 2 model (Anil et al., 2023) (proxy for quality), and embedding distance on mBERT sentence embeddings (proxy for diversity). We apply both methods to the English side of the parallel data only, to simulate the case where we don't yet have LRL translations. As a baseline, we randomly select sentences.

Table 1 shows results after finetuning. Greedy token set-cover performs the best, with diversity-based metrics actively hurting performance.

Researcher in the Loop (RITL) Despite its success in the ablation, Greedy Token Set Cover had several problems when we scaled it to select from among all the English sentences of CommonCrawl. Firstly, it is maximized by honeypots, or nonsense strings dense in content words (Appendix Table B.1); and secondly, it biases towards short sentences, causing length distribution artifacts.

These problems are not easy to solve with heuristics—for example, if you disqualify lists with commas you'll get ones with spaces, if you require sentences to have some function words or token-length diversity, you'll get other sorts of garbled sentences, and so on. However, a dataset like SMOL is small enough to manually inspect. Therefore we develop *Researcher in the Loop Greedy Set-Cover* (Algorithm 1), where the domain expert (the researcher) can look at and edit each individual sentence.[9] The result of this process is SMOLSENT, a set which uses 863 sentences to cover 5519 unique tokens. Qualitatively, SMOLSENT consists of complex sentences with wide vocabulary coverage; quantitative metrics are explored in Appendix B.3.

Algorithm 1 Researcher in Loop Greedy Set Cover

```
Res ← ...          ▷ Sentence reservoir, e.g. CommonCrawl
Toks ← ...              ▷ Tokens to Cover, e.g. GATITOS
Cov ← {}        ▷ Set-cover, aka output of this algorithm
while not ToCover.empty() do
    batch ← TopScoringSentences(Res, Toks)
    chosen ← ResearchersChoice(batch)
    chosen ← LetResearcherEdit(chosen)
    Cov.add(chosen)
    RemoveCoveredToks(Toks , chosen)
    Res ← LetResearcherDiscardSentences(Res)
    Res.remove(chosen)
end while
return Cov
```

3.2 SMOLDOC: LLMs with prompt mesh

SMOLDOC follows a different and complementary approach. Whereas SMOLSENT consists of a small set of *sentences* that are *selected* from natural text, are *complex*, and cover many *tokens*; SMOLDOC instead consists of *documents* that are *generated* and are *simpler*, but cover many *topics*. It should be noted that the token-coverage approach described above failed resoundingly for longer documents, as the prevalence of the honeypots was magnified.

To generate SMOLDOC, we used a collection of templates to create a few thousand diverse prompts with a wide range of topics, domains, words, tenses, grammatical cases, and registers (e.g. formal, informal). Appendix C.2 gives details and examples.

Corpus Diversity Ranking for SMOLDOC Document-by-document evaluation as described above does not help one understand *corpus diversity*—for example, if an almost identical document appears twice, only one of them should be included.

[9]This work was conducted before the advent of LLMs. Today, this could be simplified using LLMs as autoraters.

Therefore, we rank all candidates by how much new information they add to the corpus, by iteratively finding the document contributing the least new information and removing it, thus ranking all documents. Our criterion for "new information" was the average character 9-gram Inverse Document Frequency (IDF) score of a document—in other words, how rare its substrings were across all of the documents in the pool so far. To down-weight internally repetitive documents, we substracted the fourth moment BREAD score (Caswell et al., 2023).

Language Tiers for SMOLDOC We wanted to translate more data for languages with more speakers. We break the languages into the five different groups, each with a larger subset of the generated documents. Each tier contains translations of the top N documents as ranked by corpus diversity. These can be seen in Appendix Table C.1.

Non-English-centric translations For SMOLDOC, we additionally collected data for four non-English-centric language pairs, from each of the East African languages of Amharic (`am`) and Swahili (`sw`) to each of regionally relevant languages Standard Arabic (`ar`) and Mandarin Chinese (`zh`). Including the reversed versions of these, this yields 8 total language pairs. Because of the difficulty of generating good source material in these languages, we used the existing SMOLDOC translations to Swahili/Amharic as the source text. However, due to the lack of appropriate evaluation sets, it is hard to know the value-add of this data over datasets pivoted through English.

4 Data Collection and Verification

Several languages are contributed by volunteers; they are listed as co-authors.[10] For the other languages, the translation provider we contracted has worked with us for many years, and has a pre-existing relationship with professional translators for all languages in the SMOL datasets. The translators are paid a fair wage, and their identities are contractually kept anonymous to us. We checked the delivery for duplicate translations, anomalous source/target length ratios, and similarity with Google Translate outputs. Very few languages were flagged this way. Following this, we ran FUNLANGID (Caswell, 2024) on all segments and found no issues. Manual inspection turned up several issues with nonunicode fonts (e.g. ô for ɔ) for West African languages, and nonstandard orthography for Santali; these issues were then fixed. The choice of script, orthography and translation variety was challenging for many communities, including Kurdish, Zaza-Gorani and Gilaki languages, all of which have more than one orthography and lack a standard variety.

The largest missing check is for fluency, which is hard to measure without trusted native speakers *outside of the translation agency*, or trusted LLMs; neither of which exist for all SMOL languages.

5 Finetuning and In-Context Learning

We use fine-tuning and ICL as tools to demonstrate the value of the SMOL dataset. As this is a data paper, these experiments are motivated by the maxim *"what could any researcher simply train with public APIs?"* More involved techniques, e.g. Reinforcement-Learning-based approaches, will likely lead to stronger results.

5.1 Evaluation

Since so many language pairs are covered, we evaluate on a combination of all available evaluation sets, namely FLORES-200 (Team et al., 2022), NTREX (Federmann et al., 2022b; Barrault et al., 2019), and an in-house eval set. Since no reliable embedding models exist for these languages, trained metrics are not an option, so we use CHRF (Popović, 2015) as implemented in SacreBleu (Post, 2018)[11] with NFKC unicode normalization as our metric. For ten-shot decoding, exemplars were selected from both sub-datasets of SMOL using CHRF-counterweighted RAG (Appendix D).

5.2 Finetuning Setup and Results

We finetuned Gemini 2.0 Flash for 40 epochs on SMOLDOC, SMOLSENT, a combination of the two (BOTH), and their combination plus GATITOS (BOTH+G). To simplify finetuning, we split SMOLDOC into sentence pairs (SMOLDOCSPLIT).

Results can be seen in Table 3. Finetuning on SMOLSENT gives an average gain of +2.7 CHRF points, and SMOLDOCSPLIT gives +2.6 CHRF points on its languages. Concatenating the two training datasets leads to a gain of +3.3 to +3.6 CHRF points, and adding in GATITOS bumps it

[10] Community contributions of translations or corrections are welcome; please reach out to the authors or join the TUSL Discord.

[11] signature: `case.mixed+numchars.6+numrefs.1 +space.False+tok.none+version.1.3.0`

	Total Dataset				Per Language Pair (LP)		
Set	# languages	Examples	Tokens	Characters	Examples	Tokens	Characters
GATITOS	176	693k	784k	4.6M	3.9k	4.5k	26k
SMOLSENT	81	70k	994k	6.1M	863	12k	75k
SMOLDOC	100	27k	5.1M	28M	263	50k	278k
BOTH	115	97k	6.1M	34M	827	52k	294k

Table 2: Statistics for the whole data set (left bloc) and per language-pair (LP) (right bloc) on the two SMOL datasets and their predecessor GATITOS in number of examples, tokens, and characters. The `# languages` column counts translated languages only, not the source languages of English, Swahili, and Amharic.

LP subset →	SMOL-SENT (80 LP)		SMOL-DOC (73 LP)		Intersect (38 LP)		HARD (32 LP)	
Model ↓	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot	0-shot	10-shot
G. TRANSLATE	-	-	-	-	**43.2**	-	-	-
NLLB-54B	-	-	-	-	40.0	-	-	-
CLAUDE 3.5 SON.	37.5	**39.7**	38.3	**40.9**	41.0	42.8	30.0	**33.5**
GPT-4O	29.9	34.1	31.8	36.3	35.4	38.5	15.9	23.7
GEMINI 2.0 PRO	**38.9**	38.9	**39.9**	40.3	42.6	42.2	**31.4**	31.7
GEMINI-2.0 FLASH	35.6	38.4	36.9	39.7	40.2	41.4	26.3	30.4
+ SMOLSENT	38.3	38.3	38.8	38.8	40.6	40.6	32.5	32.6
+ SMOLDOC	35.3	35.4	39.5	39.5	41.2	41.2	31.8	31.8
+ BOTH	38.9	38.9	40.5	40.5	41.8	41.8	33.4	33.4
+ BOTH+G	**39.4**	**39.3**	**41.0**	**40.9**	**42.1**	**42.2**	**33.9**	**33.9**
Δ_{FT}	+3.8	+0.9	+4.1	+1.2	+1.9	+0.8	+7.6	+3.5

Table 3: Finetuning Gemini 2.0 Flash on SMOL for four subsets of language pairs. The first two columns show LPs in SMOLSENT and those in SMOLDOC, to show the different effects of each split. The third shows those in both SMOL datasets AND the closed domain NMT models, for an even comparison to NMT models. Finally, the HARD column shows LPs in both SMOL splits but NOT in Google Translate, or not closely related to a language in Google Translate, to approximate the especially hard languages to learn.

up to +3.8 to +4.1 CHRF points, passing all baselines except Google Translate. The 10-shot RAG results on the un-tuned model are very close to the finetuned 0-shot results, and the finetuned models show no benefit from multi-shot decoding, suggesting that these are two different ways of giving the same information—inference-time versus training time. The 10-shot random results (not included in table) were much lower.

Gains were highest on languages that are not related to mid- or high-resource languages, and lowest on dialects close to major languages. As a heuristic to measure this, we exclude languages that are on Google Translate or closely related to languages on it (Appendix G). The average gain on these languages jumps to +7.6 CHRF.

Figure 1 shows the learning curve on a development subset of 37 languages. Although it may be surprising that so many epochs are needed before convergence, we found that further increasing learning rate led to overfitting. The sharp drop near the beginning suggests a domain mismatch between pretraining and finetuning, and suggests that the same data could be used much more effectively with a better training set-up than explored here.

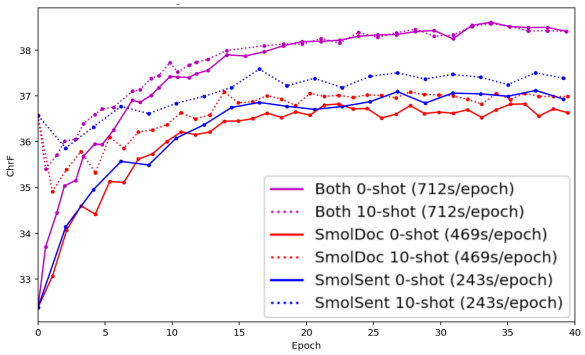


Figure 1: Training curves (CHRF) for finetuned models on a subset of 37 en→xx language pairs, trained on SMOLDOC, SMOLSENT, and their combination BOTH.

5.3 The Problem with xx→en training

Our initial experiments used all data for both en→xx and xx→en . However, the models lost performance on all tasks. The root cause turned out to be the multiway-parallel data with English tar-

Rating	Definition of Rating
N/A	True/False does not apply here. Most stories, dialogues, or fictional works would be considered N/A, unless they are promoting a falsehood about the real world.
Not Sure	Claims are made that may not be true, but you aren't sure. Choose this if it would take over 10 minutes to verify the factuality of the claim.
No Issues	All claims are factual and accurate. (Out-of-date is fine, e.g. "Barack Obama is the US President")
Minor Issue(s)	There are small inaccuracies. E.g., it may be broadly correct but frame something in a misleading way.
Clear Issue(s)	There are clear mistakes in factuality.

Table 4: Factuality Rubric

gets. LLMs are especially susceptible to repetition in data (Lee et al., 2022), and with 115 language pairs, for every one epoch over the data, the model saw about 115 epochs for each individual target sentence. Therefore, it wildly overfit and lost performance on all language pairs. Mitigating such overfitting is an important research direction to pursue, since many promising datasets are multiway parallel, e.g. FLORES-101 (Goyal et al., 2022), FLORES-200 (Team et al., 2022), NTREX (Federmann et al., 2022b; Barrault et al., 2019), and others. However, this is out of scope for the present paper, so we restrict our experiments to en→xx .

Seeing the same *source* many times likely also has deleterious effects and should also be studied; but these effects, if they exist, are small enough that we were still able to see net gains.

6 Factuality Review

Since SMOLDOC contains LLM-generated sources, they contain some factual inaccuracies. We therefore do a full human audit and assign factuality codes to each document. Each of the 584 documents was rated by three raters. Each rating is accompanied by a detailed explanation, including sources cited. Inter-annotator agreement was high, with Cohen's κ between each pair of raters between 0.82-0.87. The error code distribution can be seen in Figure 2. The rubric is presented in Table 4.

All ratings and rationales are made available. In addition, each datum in SMOLDOC is given with a simple `factuality` annotation, which has the value `has_errors` if any one of the ratings was any of `Minor Issues` or `Clear Issues`, and `ok` otherwise. For some use-cases, like question-answering, practitioners may want to filter out nonfactual data; for others, like translation, one may not be troubled by factual errors. In addition to filtering, this also provides the first factuality dataset for most of these languages.

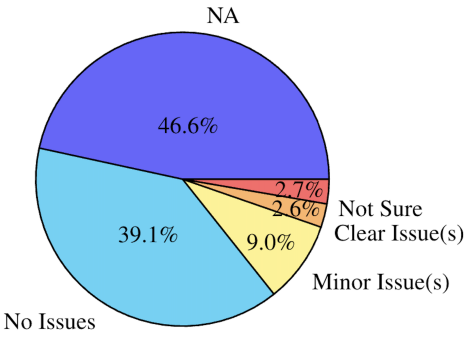


Figure 2: SMOLDOC factuality ratings.

7 Conclusion

We have open-sourced the SMOL dataset, a professionally-translated dataset covering 123 low resource languages and targeting the tasks of translation and factuality. It comprises SMOLDOC and SMOLSENT, two training datasets with the complementary strengths of sentence selection (complex, and high token coverage) and document generation (contextual, varied domains, simpler sentences) respectively. We demonstrate that finetuning Gemini 2.0 Flash on these yields to substantial improvements in translation quality. SMOL joins a growing body of resources to support underserved languages in the age of AI.

8 Limitations

The SMOL data would benefit from a more thorough review, audit, and correction from community members outside of the translators who created it. Future work on SMOL-like datasets should also focus on non-English source text that is not only maximally authentic in the given language, but also covers the topics and concepts most relevant to those languages. This approach is more difficult and would require significant work and review to do correctly. Finally, more research is needed to understand and prevent the overfitting that comes with multi-way parallel data.

References

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021a. MENYO-20k: A multi-domain English-Yorùbá corpus for machine translation and domain adaptation. *CoRR, arXiv:2103.08647v1*.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021b. "The Effect of Domain and Diacritics in Yoruba–English Neural Machine Translation". In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

Sina Ahmadi, Rico Sennrich, Erfan Karami, Ako Marani, Parviz Fekrazad, Gholamreza Akbarzadeh Baghban, Hanah Hadi, Semko Heidari, Mahîr Dogan, Pedram Asadi, Dashne Bashir, Mohammad Amin Ghodrati, Kourosh Amini, Zeynab Ashourinezhad, Mana Baladi, Farshid Ezzati, Alireza Ghasemifar, Daryoush Hosseinpour, Behrooz Abbaszadeh, Amin Hassanpour, Bahaddin Jalal Hamaamin, Saya Kamal Hama, Ardeshir Mousavi, Sarko Nazir Hussein, Isar Nejadgholi, Mehmet Ölmez, Horam Osmanpour, Rashid Roshan Ramezani, Aryan Sediq Aziz, Ali Salehi Sheikhalikelayeh, Mohammadreza Yadegari, Kewyar Yadegari, and Sedighe Zamani Roodsari. 2025. "PARME: Parallel Corpora for Low-Resourced Middle Eastern Languages". In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30032–30053, Vienna, Austria. Association for Computational Linguistics.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. Machine Translation For African Languages: Community Creation Of Datasets And Models In Uganda. In *3rd Workshop on African Natural Language Processing*.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique. *CoRR*, abs/2104.05753.

Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. 2011. "Multi-Strategy Approaches to Active Learning for Statistical Machine Translation". In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021a. NLP for Ghanaian Languages. *CoRR*, abs/2103.15475.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh,

Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021b. English-Twi Parallel Corpus for Machine Translation. *CoRR*, abs/2103.15625.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. 2022. Building Text and Speech Datasets for Low Resourced Languages: A Case of Languages in East Africa. In *3rd Workshop on African Natural Language Processing*.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building Machine Translation Systems for the Next Thousand Languages.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. "Findings of the 2019 Conference on Machine Translation (WMT19)". In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. "No Data to Crawl? Monolingual Corpus Creation from PDF Files of Truly low-Resource Languages in Peru". In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Isaac Caswell. 2024. FunLangID. `https://github.com/google-research/url-nlp/tree/main/fun_langid`.

Isaac Caswell, Lisa Wang, and Isabel Papadimitriou. 2023. Separating the wheat from the chaff with BREAD: An open-source benchmark and metrics to detect redundancy in text. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 324–338, Singapore. Association for Computational Linguistics.

Moussa Koulako Bala Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. Machine Translation for Nko: Tools, Corpora and Baseline Results.

Roald Eiselen and Martin Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. *CoRR*, abs/2004.00648.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022a. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022b. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel cor-

pus for amharic-english machine translation. *CoRR*, abs/2104.03543.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Gilles Hacheme. 2021. English2gbe: A multilingual machine translation model for {Fon/Ewe} gbe. *arXiv preprint arXiv:2112.11482*.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Asmelash Teka Hadgu, Gebrekirstos G. Gebremeskel, and Abel Aregawi. 2022. HornMT. https://github.com/asmelashteka/HornMT.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. "GATITOS: Using a New Multilingual Lexicon for Low-resource Machine Translation". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 89–95, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. "Deduplicating Training Data Makes Language Models Better". In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka English - isiZulu Parallel Corpus. Thank you to Facebook Research for funding the creation of this dataset.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. "Investigating an Approach for Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi". In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20, Marseille, France. European Language Resources Association (ELRA).

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr, et al. 2021a. A large-scale study of machine translation in the Turkic languages. *arXiv preprint arXiv:2109.04593*.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Behzod Moydinboyev, Sar-

dana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, et al. 2021b. Evaluating multiway multilingual NMT in the Turkic languages. *arXiv preprint arXiv:2109.06262*.

Wilhelmina Nekoto, Julia Kreutzer, Jenalea Rajab, Millicent Ochieng, and Jade Abbott. 2022. Participatory Translations of Oshiwambo: Towards Sustainable Culture Preservation with Language Technology. In *3rd Workshop on African Natural Language Processing*.

Elizabeth Nielsen, Isaac Rayburn Caswell, Jiaming Luo, and Colin Cherry. 2025. Alligators all around: Mitigating lexical confusion in low-resource machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 206–221, Albuquerque, New Mexico. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores.

Allahsera Auguste Tapo, Michael Leventhal, Sarah Luger, Christopher M. Homan, and Marcos Zampieri. 2021. Domain-specific MT for Low-resource Languages: The case of Bambara-French. *CoRR*, abs/2104.00041.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zaitinkhuma Thihlum, Vanlalmuansangi Khenglawt, and Somen Debnath. 2020. Machine Translation of English Language to Mizo Language. In *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pages 92–97.

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. *CoRR*, abs/2010.06354.

A Operational Definition of LRL

In this paper, we operationally define LRL as any language beyond the first 100 supported by most traditional crawls and MT providers. Since there is some variation in which languages exactly this is, we concretize it as the set of 104 languages supported in Google Translate prior to 2020. These are the languages for which launchable quality was possible before LLM-type models like M4 (Arivazhagan et al., 2019; Bapna et al., 2022) and PaLM (Anil et al., 2023) came on these scene. It is also worth noting that since these languages were on the product for much longer, they have much more machine-translated content online from services that used Google Translate for internationalization. These 104 languages are: `af am ar az be bg bn bs ca ceb co cs cy da de el en eo es et eu fa fi fil fr fy ga gd gl gu ha haw hi hmn hr ht hu hy id ig is it he ja jv ka kk km kn ko ku ky la lb lo lt lv mg mi mk ml mn mr ms mt my ne nl no ny pa pl ps pt ro ru sd si sk sl sm sn so sq sr st su sv sw ta te tg th tr uk ur uz vi xh yi yo zh zh-Hant zu`.

Rightly speaking, the languages outside of this set might better be termed "Very Low-Resource" instead of just "Low Resource", since the 104 languages above do include languages like Hawaiian, Javanese, Yiddish, and Hmong, which are by no stretch of the imagination high-resource. We will leave more rigorous definitions to future work.

B SMOLSENT details

B.1 Evaluating the SMOLSENT selection process

In Section 3.1, we describe experiments used to validate the selection process for SMOLSENT. We train a backbone MT system is pretrained on the MADLAD-400 dataset (Kudugunta et al., 2023). The following languages are held out of the training data to be used for fine-tuning experiments: Catalan, Icelandic, Marathi, Turkish, Maltese, Xhosa, Tamil, Basque, and Tajik. The model itself is a 1B parameter encoder-decoder Transformer and is trained from scratch on the MADLAD-400 data. Each of the candidate data selection methods is used to select data from the held-out languages, and then each candidate set is used in turn to finetune the backbone model. The results of this finetune step are reported in Table 1, where the set-cover approach is shown to be most effective.

B.2 Notes on Researcher in the Loop

Researcher In the Loop extends the greedy set cover approach thusly: rather than always picking the highest-scoring sentence, we iteratively show the researcher a batch of the 20 highest scoring sentences according to several scores, and let the researcher pick and optionally edit each sentence at each iteration. At each iteration, the researcher may also remove any number of this batch's sentences from the reservoir. Allowing the researcher to see and edit the sentences allows ensures that the sentences are of high-quality. To deal with the length bias issue, we showed not only sentences that maximize coverage percent, but also that maximize heuristics that weighted the coverage with the number of new tokens hit, like `log(coverage_percent)*n_hits` .

As described in the paper text, this approach is designed to combat issues such as honeypot sentences. Example Honeypot sentences can be seen in Table B.1

Sentence
Individual determine can get prolonged, reduce along with attractive.
Sell hand mood situation connect proper decision today spread true.
Demand indeed off forget act special well treat sometimes notice.
Agree board book oh trust by attractive supply deal together.
Picture exactly could ability impact advance then same admire across.
One physically courage both information language issue laugh common.

Table B.1: Honeypot Sentences for Greedy Selection: CommonCrawl has many sentences packed with content words but with no clear semantics or grammar.

B.3 Corpus statistics on SMOLSENT

To measure "Bang for our Buck" we define the *excess-token ratio* ξ as the number distinct tokens in the set cover divided by the number of target tokens, and use it along with the coverage percent to understand the SMOLSENT dataset. Table B.2 compares corpus statistics of SMOLSENT to four other corpora. `sametoks` picks a random set of sentences from CommonCrawl until it has the same number of tokens as SMOLSENT; this only covers 50% of the target tokens and has an excess-token-ratio ξ ratio of 3.3, much worse than SMOLSENT's value of 2.3. The `samecov` baseline randomly picks common-crawl sentences until it has the same token coverage as SMOLSENT, which necessitates

set	N sent	toks	types	$\xi(\downarrow)$	cov%(↑)
SMOLSENT	863	12k	5.5k	2.3	99.6%
sametoks	863	12k	3.8k	3.3	50.4%
samecov	57578	877k	38k	23.1	99.6%
SMOLDOC-T1	6979	108k	8.8k	12.3	80.5%
SMOLDOC-T5	820	12k	2.8k	4.3	40.2%

Table B.2: Corpus statistics of SMOLSENT, random selections of sentences from CommonCrawl, and tiers 1 and 5 of SMOLDOC.

set	N langs	ex	tok	char	ex/LP	tok/LP	char/LP
SMOLDOC-t1	5	2.9k	537k	3.0M	584	107k	604k
SMOLDOC-t2	31	14k	2.7M	15M	450	88k	495k
SMOLDOC-t3	24	6.7k	1.2M	6.8M	280	50k	281k
SMOLDOC-t4	8	1.0k	184k	1.0M	126	23k	128k
SMOLDOC-t5	34	2.2k	401k	2.2M	66	12k	65k
all-SMOL	GTr	97k	6.1M	34M	827	52k	294k

Table C.1: Statistics on the languages in the individual tiers of SMOLDOC.

a 67x larger set of sentences, and a correspondingly bloated excess token coverage ratio of 23.1. As a further reference we compare tiers 1 (largest) and 5 (smallest; comparative size to SMOLSENT) of SMOLDOC. As expected of machine-generated text, they have a worse ξ value, corresponding to a narrower spectrum of vocabulary used.

C SMOLDOCDetails

C.1 SMOLDOC data tier details

Per-tier statistics on the SMOLDOC dataset can be seen in Table C.1.

C.2 Details on SMOLDOC prompt creation

To avoid biases from overly tempted prompts, we put in significant effort to make sure the prompts were all very different. Each prompt drew at random from the following elements:

- random selection of English words to use in the response
- one of 600 manually created topics, e.g. “volcanic eruptions” or "A special tree”
- one of 50 tone/tense categories, e.g. “Please use the subjunctive mood.”, “Use an effusive tone.”
- A style prompt, e.g. “You are the author R.K. Narayan.” or “You are a mother talking to her son.”
- A text modality, e.g. story/dialogue/essay

In addition to this, we added a few more sources of prompts:

- Prompts based on urls, meant to simulate different web domains, like Wikipedia and reddit.
- Prompts based on continuing the sentences from SMOLSENT
- Prompts based on current events, history, and daily life in different countries
- Special effort was made to include dialogues (to get more spoken register) and recipes (unique domain that may also be important to translate).

For each prompt, we generated 8 responses (T=0.7). These were ranked by their simple token density (unique tokens over total tokens), and the top two were chosen for consideration. Using the Researcher-in-the-loop mentality (“measure twice cut once!”), we went over 1000+ responses by hand and scored/edited them. This was mainly to filter out questionable or boring responses. A typical paragraph scored as 0 would be LLM-speak like *“X is a complex and multifaceted problem with no easy solution. Here are some suggestions. Keep in mind that there is no one-size-fits all solution, and ultimately, the choice is up to you. [...]”*.

Example prompts can be seen in Table C.2.

C.3 SMOLDOC Errata

Orthographies Several languages use irregular orthographies. Most notable is Mooré/Mossi (mos), where different translators have used a variety of different conventions. After soliciting community feedback, we plan to release standardized versions to the data.

Example Prompts
You are Ernest Hemingway. Write a dialogue about road rage. Use a didactic tone.
Write a 1-paragraph story concerning an Irish wake.
You are a teenager talking to his friend. Please carefully craft a 1-paragraph bit about an engineer who subsists off coffee. Try to include the words “confirmed”, “move” and “above”.
Give a typical, yet interesting, example of something you would find on reddit.
Please write a few paragraphs about challenges facing Ethiopia.
Please write a long passage starting with ‘Mum and Dad pause their debate when we hear this creepy clacking that sounds like hail falling.’
Write a recipe for baking an almond cake.

Table C.2: Representative sample of prompts use to generate the documents for SMOLDOC

document selection When collecting the data for SMOLDOC for the Indian languages, we mistakenly included a variety of documents that fell below the corpus diversity threshold described in Section 3.2.

D N-shot: CHRF-counterweighted RAG

To have a strong baseline for N-shot results, we adopt a RAG-based approach that resembles the greedy set-cover algorithm. For each sentence in the eval set, we want the best coverage of the source sentence n-grams as possible, with the least redundancy among exemplars. Therefore, we iteratively choose the exemplar whose source side has the minimum CHRF to the eval source. However, when counting the true positives in the CHRF calculation, we weight the count of each ngram n_i by $(1+c_i)^{-\alpha}$, where $c_i \in [0, \infty]$ is the number of times n_i has been seen among the exemplars chosen so far, and α is a parameter to control how close this algorithm is to ngram set-cover. We use $\alpha = 2$.The set of exemplars we choose from is the concatenation of SMOLSENT and SMOLDOCSPLIT.

E Prompts for Decoding

For 0-shot prompting, we used the following, fairly wordy prompt, the SL and TL standing for the source and target language name, respectively:

```
You are an expert translator. I am
going to give you some example
pairs of text snippets where the
first is in ${SL} and
the second is a translation of
the first snippet into
${TL}. The sentences
will be written
${SL}: <first sentence>
${TL}: <translated first sentence>
After the example pairs, I am
going to provide another sentence
in ${SL} and I want
you to translate it into
${TL}. Give only the
translation, and no extra
commentary, formatting, or
chattiness. Translate the text
from ${SL} to ${TL}.
```

For finetuned models, there is no need for such a wordy prompt, and indeed it only risks overfitting. Therefore, we used the following minimalist prompt:

```
Translate from ${SL} to ${TL}:
```

F Volunteer contributions

A few languages have extra details that need to be called out here.

F.1 Translations for Cantonese

A volunteer team of Cantonese speakers at Google pulled together to translate the maximal set of SMOL text. Mingfei Lau and Jonathan Eng were the main leaders of this effort, and the contributors to translation and post-editing were (alphabetically): Tsz Yan Au, Emily Awesome, Jason Chan, Siu Man Chan, Vicky Chan, Yiwang Chen, Kinton Cheung, Mingo Choi, Andy Chow, Ashley Chow, Olivia Chow, Daniel (Ying Wai) Fan, Thomas Fung, Vikki Ha, Joshua Kwong, Liam Lee Pong Lam, Jonas Lau, Ying Tung (Grace) Law, Crystal Lee, Aki Leung, Derek Leung, Jackie Leung, Thomas Leung, Mu Li, Alicia Liu, Malena Loosli, Chui McConnell, Ken Ng, Nicholas Ng, Tonia Shen, Helen Shum, Franky Sze, Eric Tang, Tommy Tse, Daniel Wong, Danny Wong, Maggie Wong, Pinki Wong, Jeffrey Yu, Shanelle Yu, Shing Fung Yue, Miranda Zhang, and Willis Zhang.

F.2 Translations for NKo

The initial delivery for the NKo language (nqo) had a wide variety of errors. We reached out to the authors from Doumbouya et al. (2023), who did a complete re-translation of the text.

F.3 Translations for Zazaki, Hawrami, and Gilaki

Sina Ahmadi gratefully acknowledges support from the UZH Postdoc Grant (reference number 269093).

F.4 Translations for Zarma

Annotation Pipeline The Zarma translation process of SMOL—all the subsets—was done through a combination of automatic and human in the loop methods. We leveraged some existing tools that our team developed to speed up the annotation process. We first used a baseline bidirectional model that we developed to produce initial translation of the samples. These machine translated samples were then passed through our Zarma grammatical error correction model. This model was built by pre-training gemma-2-9b on Zarma data and fine tuning the checkpoint on grammar error correction data set using Direct Preference Optimization (DPO) settings. The outputs from this stage—both languages side by side—were then given to our team of annotators for review.

The annotators were given some guidelines—in addition to the general guidelines from SMOL—for the annotations. These guidelines include:

- **Word adaptation:** rules for handling technical terms, proper nouns, and domain-specific vocabulary that might not have direct equivalents in Zarma. E.g: all the scientific/technical words remain unchanged; and words that have known french-ized equivalent in Zarma must be used in their french-ized forms (for better understandability).

- **Prioritize understandability:** guidelines to prioritize understandability and fidelity over word-for-word translation. We instructed annotators to focus on creating translations that sound natural and widely understandable by Zarma speakers.

- **Language specific constraints:** language specific guidelines that cannot be generalized.

The pipeline speeds up the process while maintaining the quality, since some of the outputs from the automatic stages were already correct.

Zarma Community Attitude Towards Tech : The Zarma community—and the whole Niger in general—are very open minded regarding technology. When we started our very first resource creation for the Zarma language, we received positive feedback and even help from the community, as long as we developed an openly accessible solution for the community. **For the SMOL annotation, that trust helps us to receive valuable help.** For instance, a government based institution verbally promised to accompany any language preservation—machine learning focus in our case—if the outcome will be open-sourced for community usage.

F.5 Post-Edits for Mooré

Annotation Pipeline The annotation process for Mooré did not involve any automated components; everything was annotated by humans. The annotation focuses more on the guidelines provided by SMOL, in addition to some more as in the Zarma case.

Mooré Community Perspective The Mooré community, similarly to the zarma community, are very open minded towards technology; especially if it touches cultural/language preservation. One main feedback we received from some elders (parents of one of the annotators) was a warning to ONLY USE standard Mooré orthography, not any equivalent. They want the language to be well documented according to the language standards.

F.6 Post-Edits for Indonesian

A volunteer team post-edited translations of smoldoc and smolsent datasets that had done by Gemini 2.5 Pro. The contributors to translation and post-editing were Muhammad Ravi Shulthan Habibi, David Anugraha, and Genta Indra Winata. The post-edits resulted in about 70% of the machine translations being changed.

Translators agreed that the system output was often too “formal", “stiff", or “awkward". The “formal" translations were furthermore not formal in an acceptible sense, but “too awkward and stiff, even for a more formal situations", as an annotator said. Each word choice might be correct and standard in Indonesian, but when combined in a sentence, the result sounded unnatural. Therefore, the majority

of the post-edits focused on making the translations sound more natural.

Nonetheless, overall the system output was already quite reasonable in terms of register. In some cases, though, it leaned toward being too rigid. The post-edits tried to loosen that into a consistent "medium" range, but with some flexibility depending on the style of each sentence (sometimes slightly more formal, sometimes slightly less) so the overall text still feels natural and coherent.

F.7 Translations for Languages of the Russian Federation

Traditionally, speakers of hundreds of Cyrillic-based languages in the Russian Federation translate datasets via Russian. For the success of this project, I (Ali Kuzhuget) first funded a professional translation into Russian, engaging Andrey Anisimov as the main translator. The proofreading was conducted by Farhad Fatkullin, Vice-President of the National League of Translators, together with machine translation specialist David Dalé. I also oversaw formatting correctness and coordinated the overall translation workflow.

In parallel, I supervise the translation of the dataset from Russian into Tuvan, using a dedicated Telegram chatbot for large-scale dataset translation. This tool enables multiple rounds of validation and systematic assessment of translation quality. Currently, representatives of about a dozen Cyrillic languages are in the process of translating the SMOL dataset into their own languages through Russian and/or English (for example, Tuvan, Bashkir, Chuvash, and others), ensuring both linguistic accuracy and cultural relevance.

G Full results

Full per-language results can be seen in Table G.1. Results are sorted by the Δ_{FT}, which is the CHRF of the BOTH model minus the CHRF of the finetuned BOTH model—in other words, how much the finetuning on SMOL improved the baseline model.

Google Translate Languages and their cousins As mentioned in the results section, some languages see only very small improvements from finetuning on SMOL, and others even see losses. These are mainly either high-resource languages, or close relatives to higher-resource languages. In the full table G.1 below, The superscript $^{\text{GTr}}$ indicates a language supported by Google Translate at the time of these experiments; a superscript like $^{\sim\text{xx}}$ means that this language is closely related to the Google-Translate-supported language xx. We only consider the 108 languages that were present on Google Translate at the time of this work.

lang	cat	Δ_{FT}	G2F	+sS	+sD	+sB	+sG	Cld	+RAG	G2P	GPT4o	GTr	NLLB
ee	BOTH	+36.1	3.0	37.7	37.6	39.1	39.2	37.8	40.0	39.5	7.5	**42.7**	40.7
kr	BOTH	+10.8	17.3	25.6	25.9	28.1	28.8	22.7	26.3	20.3	22.2	**32.6**	31.0
kg	BOTH	+9.2	34.9	46.9	36.8	44.1	43.2	43.2	47.0	37.8	29.1	**50.2**	3.4
bem	BOTH	+7.3	40.0	44.8	44.7	47.3	49.2	43.3	47.7	42.3	33.3	**49.7**	41.8
dyu	BOTH	+5.3	17.9	22.5	23.3	23.2	23.7	23.9	**24.4**	21.0	4.5	22.4	12.5
din	BOTH	+4.6	20.3	23.8	22.9	24.9	25.1	23.3	25.9	21.4	1.6	25.1	**26.5**
luo	BOTH	+4.1	37.4	39.1	41.1	41.5	42.0	39.1	**42.0**	39.6	36.1	41.3	39.5
fon	BOTH	+3.6	21.3	24.3	23.9	24.9	25.3	20.4	23.7	23.8	1.9	**25.9**	24.2
bm	BOTH	+3.4	30.8	28.6	35.2	34.2	34.1	34.0	**36.2**	33.9	9.0	35.7	32.2
ak	BOTH	+2.6	35.5	36.1	37.8	38.1	**38.2**	34.4	38.1	37.3	32.2	34.5	33.3
ln	BOTH	+2.5	46.8	48.1	48.4	49.3	**49.3**	44.6	48.3	46.5	45.2	46.4	45.7
wo	BOTH	+1.1	30.3	30.0	30.4	31.4	31.6	31.4	32.2	30.7	29.8	**36.2**	30.9
ff	BOTH	+0.9	25.0	24.4	26.4	25.9	26.5	25.7	26.1	25.2	2.5	25.9	**27.1**
om	BOTH	-0.8	40.1	38.0	39.0	39.3	39.4	39.0	40.2	41.3	38.4	**41.4**	39.1
lg	BOTH	-1.1	42.5	39.9	41.4	41.4	41.7	42.0	43.1	43.5	41.0	**43.6**	41.1
ber	BOTH	-3.2	25.3	20.6	22.9	22.1	21.9	28.5	25.2	31.1	2.8	21.0	**32.4**
trp	SMOLDOC	+29.7	8.4	6.5	37.8	38.1	**39.1**	24.7	35.8	27.2	20.3	35.9	-
mni-M.	SMOLSENT	+26.4	2.9	30.0	1.2	29.3	29.3	29.6	31.8	33.6	1.3	**45.6**	0.8
gaa	BOTH	+23.1	22.7	44.5	44.4	45.8	47.4	34.7	44.0	40.9	6.6	**48.3**	-
dov	BOTH	+21.1	19.1	39.2	38.3	40.2	40.6	19.2	39.5	18.2	8.7	**41.7**	-
ahr$^{\sim hi}$	neither	+17.8	24.2	31.8	41.9	42.0	**42.8**	32.8	39.0	30.0	36.9	-	-
sus	BOTH	+17.8	11.3	28.3	26.8	29.1	30.3	26.1	29.4	20.7	5.6	**34.6**	-
nqo	BOTH	+17.5	0.2	17.9	17.1	17.7	17.5	17.1	17.9	17.2	1.1	**19.1**	-
alz	BOTH	+15.5	16.9	31.5	30.5	32.4	33.4	25.3	30.6	26.9	8.9	**36.6**	-
lu	BOTH	+13.8	27.6	37.5	39.3	41.4	**42.2**	27.2	37.0	34.8	21.9	-	-
cgg	BOTH	+12.2	32.6	40.5	40.5	44.8	**44.8**	37.3	42.2	37.7	28.3	42.8	-
ks-D.$^{\sim hi}$	neither	+11.7	14.9	26.6	18.8	26.6	**27.7**	23.8	27.1	21.5	19.2	-	21.1
brx	SMOLSENT	+11.6	24.3	36.0	0.4	35.9	**37.0**	30.8	35.9	36.2	5.2	-	-
mag$^{\sim hi}$	neither	+8.1	47.0	45.3	55.7	55.1	54.7	47.3	51.8	47.4	48.7	-	**59.4**
ki	BOTH	+7.7	32.6	38.2	38.0	40.3	40.6	35.9	**42.0**	39.1	10.5	-	38.4
aa	BOTH	+7.4	14.2	20.1	20.3	21.6	21.8	18.9	20.6	18.9	5.6	**23.1**	-
ks$^{\sim ur}$	neither	+7.3	22.1	29.4	0.4	29.4	29.7	28.0	30.4	30.5	26.3	-	**36.7**
nr$^{\sim zu}$	neither	+7.0	48.9	54.0	51.1	55.9	57.5	48.0	53.6	51.2	45.5	**59.5**	-
doi$^{\sim hi}$	neither	+6.6	34.3	28.1	**41.4**	40.9	41.3	35.9	39.5	38.2	27.7	40.4	-
sat-L.	SMOLSENT	+6.4	12.8	19.5	15.5	19.2	20.8	**25.3**	22.5	22.7	21.3	22.7	-
mfe$^{\sim fr}$	neither	+5.3	59.5	65.4	62.6	64.8	66.9	59.6	65.0	59.8	59.5	**67.5**	-
ach	BOTH	+5.2	33.2	43.1	32.5	38.4	39.2	32.4	37.3	35.1	23.8	**43.2**	-
ayl$^{\sim ar}$	neither	+4.5	47.3	51.7	51.9	51.8	**53.9**	45.3	48.9	46.3	48.6	-	-
st^{GTr}	neither	+4.5	49.9	54.0	55.2	54.4	55.0	49.4	**57.0**	53.1	49.2	49.0	47.2
ber-L.	BOTH	+4.2	26.1	27.9	30.3	30.3	30.9	27.6	32.7	32.1	21.2	**34.7**	-
apd-S.$^{\sim ar}$	neither	+3.6	42.3	**50.2**	43.3	45.9	47.1	43.2	45.0	42.5	45.6	-	-
ve$^{\sim sn}$	neither	+3.5	47.9	50.0	48.7	51.4	52.2	50.2	53.1	52.7	43.9	**56.8**	-
kri$^{\sim en}$	neither	+2.8	31.5	34.2	31.8	34.3	34.7	34.5	33.5	30.7	34.9	**34.9**	-
tiv	BOTH	+2.5	23.8	25.7	25.8	26.3	**26.5**	22.3	24.2	24.5	1.5	25.2	-
gn	SMOLSENT	+2.3	37.4	37.8	30.4	**39.7**	38.4	36.0	38.0	36.4	35.6	38.4	38.5
mos	BOTH	+2.3	18.2	20.9	18.9	20.5	21.1	24.3	**25.0**	20.9	1.3	-	23.8
tum$^{\sim ny}$	neither	+1.1	40.8	39.5	42.4	41.9	42.8	40.0	42.7	43.8	37.7	**45.4**	36.2
ti$^{\sim am}$	neither	+0.8	24.2	24.3	24.9	25.0	25.7	25.0	**26.2**	26.1	9.3	26.1	25.5
yo^{GTr}	neither	+0.6	34.6	33.8	35.4	35.2	35.8	29.2	**36.8**	26.7	27.6	21.3	32.6
tn$^{\sim st}$	neither	+0.2	52.5	50.8	51.9	52.7	53.2	50.1	51.7	53.3	36.8	**55.6**	53.0
ar-M.$^{\sim ar}$	neither	+0.1	40.1	41.8	39.1	40.2	40.9	40.5	40.8	40.4	41.0	-	**43.0**
am^{GTr}	neither	+0.1	34.0	33.2	33.0	34.1	33.5	31.6	32.6	**35.8**	29.6	34.7	30.3
ig^{GTr}	neither	-0.1	47.2	47.1	47.0	47.1	47.8	43.9	46.2	**47.8**	46.2	47.6	46.6
so^{GTr}	neither	-0.1	49.7	46.2	50.0	49.6	49.1	48.7	49.8	50.3	**50.8**	50.6	48.6
arz$^{\sim ar}$	neither	-0.1	48.6	46.1	48.9	48.5	47.8	49.6	**50.3**	48.8	49.7	-	49.6
kl	SMOLSENT	-0.3	40.6	39.7	30.2	40.3	41.2	42.2	**43.1**	41.2	41.6	42.9	-
sa	SMOLSENT	-0.3	33.0	32.9	26.7	32.7	33.4	31.8	33.0	32.1	32.0	**35.2**	29.0
ay	SMOLSENT	-0.5	32.7	31.8	24.2	32.2	32.4	33.4	33.2	32.9	30.0	**34.7**	31.7
sn^{GTr}	neither	-0.5	50.5	48.3	50.2	50.0	50.3	46.8	48.8	**51.8**	50.3	49.2	48.2
efi	BOTH	-0.6	14.7	14.5	14.3	14.1	14.2	**15.3**	15.1	15.0	2.2	-	-
ss$^{\sim zu}$	neither	-0.6	50.2	48.8	48.2	49.6	50.3	49.6	51.2	51.8	46.0	**56.3**	48.1
yue$^{\sim zh}$	neither	-0.7	26.8	25.8	25.1	26.1	26.1	28.2	28.2	27.5	**31.6**	25.9	22.6
bci	BOTH	-0.7	23.2	22.2	21.8	22.5	22.9	17.1	20.7	27.6	1.0	**29.8**	-
ndc-Z.$^{\sim sn}$	neither	-1.0	29.2	27.9	28.9	28.2	28.3	27.6	28.0	**29.6**	28.6	29.5	-
es^{GTr}	neither	-1.1	62.4	61.3	51.6	61.3	61.3	-	-	63.0	-	**63.5**	61.8
sat	SMOLSENT	-1.3	32.4	30.9	1.0	31.1	30.8	34.7	36.0	**36.3**	1.8	35.7	-
rw^{GTr}	neither	-1.4	45.2	43.1	43.0	43.8	43.8	43.2	44.0	45.1	44.7	**48.8**	43.4
nd$^{\sim zu}$	neither	-1.4	43.9	41.5	42.6	42.5	43.2	42.3	42.9	**44.5**	43.6	-	-
sw^{GTr}	neither	-1.5	66.7	64.6	64.2	65.2	64.8	64.0	65.5	**67.2**	66.5	65.3	60.5
mg^{GTr}	neither	-1.9	52.8	48.9	51.6	50.9	51.5	52.4	52.5	**53.3**	52.2	52.6	52.1
qu	SMOLSENT	-1.9	34.7	32.8	30.4	32.8	33.0	35.3	35.1	34.0	22.0	**36.3**	27.9
zu^{GTr}	neither	-2.0	58.3	56.4	55.3	56.3	56.1	54.1	55.5	**58.5**	57.5	57.6	57.6
lus	SMOLSENT	-2.1	42.6	39.7	38.0	40.5	41.4	40.6	41.5	**43.8**	33.5	42.6	39.0
scn$^{\sim it}$	neither	-2.2	52.4	47.3	50.0	50.2	50.8	49.9	51.4	52.1	49.5	**53.3**	51.0
nso$^{\sim st}$	neither	-2.3	46.8	42.8	43.6	44.5	44.6	46.9	47.7	**48.1**	46.9	47.6	45.5
xh^{GTr}	neither	-2.3	53.9	49.7	50.7	51.6	51.8	51.6	52.3	53.9	53.7	**54.8**	51.2
ne^{GTr}	neither	-2.7	54.3	51.8	51.7	51.6	52.1	52.4	52.5	52.4	52.7	**54.9**	45.2
pa-A.$^{\sim pa}$	neither	-3.0	38.1	35.8	0.3	35.1	35.7	41.6	41.2	36.7	37.3	**43.5**	-
aeb$^{\sim ar}$	neither	-3.3	46.5	41.9	42.3	43.2	43.4	45.9	46.8	47.6	**49.2**	-	43.8
ha^{GTr}	neither	-3.4	**54.5**	50.7	49.9	51.1	51.5	50.9	51.0	54.1	53.9	53.8	53.9

lang	cat	Δ_{FT}	G2F	+sS	+sD	+sB	+sG	Cld	+RAG	G2P	GPT4o	GTr	NLLB
ts$^{\sim zu}$	neither	-3.6	50.6	47.2	46.4	47.0	48.1	49.7	50.1	51.6	49.0	**52.9**	51.3
rn$^{\sim rw}$	neither	-3.9	44.5	39.3	40.5	40.6	40.9	43.1	43.4	**46.2**	44.3	45.4	45.0
af^{GTr}	neither	-4.2	71.9	68.8	67.9	67.7	68.3	71.7	**72.5**	72.1	71.8	71.5	68.6
bo	SMOLSENT	-4.3	41.3	36.7	34.7	37.0	37.3	42.6	42.1	**43.3**	19.8	41.8	36.9
ny^{GTr}	neither	-5.1	55.0	47.5	50.5	49.9	49.7	53.0	53.1	55.3	53.9	**55.8**	50.3
pcm$^{\sim en}$	neither	-6.5	47.9	43.5	39.4	41.4	41.6	51.3	45.7	49.8	**56.0**	-	-
tcy	SMOLDOC	-6.8	34.7	22.0	28.1	27.9	28.8	28.2	29.3	36.7	21.6	**39.1**	-
ktu	BOTH	-9.4	56.6	59.3	40.4	47.2	51.3	45.8	48.4	57.8	22.3	**64.3**	-

Table G.1: Full results (0-shot) For the en→xx direction. Languages in the Intersect subset (supported by all models) are shown first, and then all other languages. The Δ_{FT} compares the base model and the model finetuned on BOTH, to give an idea of how effective the SMOL datasets are for that language. The CAT column indicates which SMOL datasets support this language. The superscript GTr indicates a language supported by Google Translate; a superscript like $\sim$ xx means that this language is closely related to that Google-Translate-supported language.

Abbreviations: This table needed some squishing to fit. Language varieties whose script/region is different from the CLDR default would have the ISO-15924 script code in the BCP-47 code, like MNI-MTEI or BER-LATN; in this table we have abbreviated them to the first letter thereof (MNI-M or BER-L). Similarly, we have abbreviated:
SMOLSENT → sS
SMOLDOC → sD
BOTH → sB
BOTH+GATITOS → sG
GEMINI 2.0-{FLASH, PRO} → G2.0-{F,P}
GOOGLE TRANSLATE → GTR.

H Complete Per-Language details: the Big-SMOL table

A summary of all SMOL language pairs and coarse-grained information about them can be seen in Table H.1. Numbers are given in terms of examples; keep in mind that a single example in SMOLDOC is a document, whereas in SMOLSENT it is a sentence.

Lang. pair	target language name	ISO 15924 Script	Continent	trg.chars	S.DOC	S.SENT
en_yo	Yoruba	Latn	Africa	780k	584	863
en_sw	Swahili	Latn	Africa	699k	584	863
en_ha	Hausa	Latn	Africa	696k	584	863
en_grt-Latn	Garo (Latin script)	Latn	Asia	591k	457	0
en_trp	Kokborok	Latn	Asia	581k	457	0
en_mg	Malagasy	Latn	Africa	580k	391	863
en_xsr-Tibt	Sherpa (Tibetan script)	Tibt	Asia	569k	457	0
en_om	Oromo	Latn	Africa	542k	391	863
en_sd-Deva	Sindhi (Devanagari script)	Deva	Asia	525k	456	0
en_ccp-Latn	Chakma (Latin script)	Latn	Asia	521k	457	0
en_spv	Sambalpuri	Orya	Asia	508k	457	0
en_doi	Dogri	Deva	Asia	503k	454	0
en_xnr	Kangri	Deva	Asia	503k	457	0
en_mjl	Mandeali	Deva	Asia	496k	457	0
en_lif-Limb	Limbu (Limbu script)	Limb	Asia	494k	457	0
en_ne	Nepali	Deva	Asia	494k	456	0
en_kru	Kurukh	Deva	Asia	492k	457	0
en_hoc-Wara	Ho (Warang Chiti script)	Wara	Asia	492k	457	0
en_bra	Braj	Deva	Asia	491k	457	0
en_bns	Bundeli	Deva	Asia	490k	456	0
en_mag	Magahi	Deva	Asia	488k	456	0
en_wbr	Wagdi	Deva	Asia	488k	455	0
en_bfy	Bagheli	Deva	Asia	487k	457	0
en_unr-Deva	Mundari (Devanagari script)	Deva	Asia	485k	457	0
en_mtr	Mewari	Deva	Asia	480k	457	0
en_tcy	Tulu	Knda	Asia	480k	451	0
en_ahr	Ahirani	Deva	Asia	479k	457	0
en_ig	Igbo	Latn	Africa	474k	391	863
en_dhd	Dhundari	Deva	Asia	465k	456	0
en_bfq	Badaga	Taml	Asia	464k	457	0
en_kfy	Kumaoni	Deva	Asia	462k	457	0
en_bgq	Bagri	Deva	Asia	462k	457	0
en_scl	Shina	Arab	Asia	460k	457	0
en_am	Amharic	Ethi	Africa	443k	584	863
en_lep	Lepcha	Lepc	Asia	441k	456	0
en_st	Sesotho	Latn	Africa	412k	260	863
en_sgj	Surgujia	Deva	Asia	395k	356	0
en_so	Somali	Latn	Africa	392k	260	862
en_ny	Chichewa	Latn	Africa	386k	260	863
en_sn	Shona	Latn	Africa	382k	260	863
en_rw	Kinyarwanda	Latn	Africa	378k	260	863
en_zu	Zulu	Latn	Africa	373k	260	863
en_lg	Luganda	Latn	Africa	369k	260	863
en_xh	Xhosa	Latn	Africa	368k	260	863
en_ln	Lingala	Latn	Africa	365k	260	863
en_noe	Nimadi	Deva	Asia	342k	315	0
en_luo	Luo	Latn	Africa	340k	260	863
en_bm	Bambara	Latn	Africa	337k	260	863
en_ak	Twi	Latn	Africa	328k	260	863
en_sjp	Surjapuri	Deva	Asia	327k	299	0
en_wo	Wolof	Latn	Africa	321k	260	863
en_ff	Fulani	Latn	Africa	320k	260	862
sw_ar	Arabic	Arab	Asia	274k	330	0
en_ar-MA	Morrocan Arabic	Arab	Africa	273k	260	863
en_arz	Egyptian Arabic	Arab	Africa	265k	260	863
am_ar	Arabic	Arab	Asia	265k	329	0
en_nso	Sepedi	Latn	Africa	243k	130	863
en_ti	Tigrinya	Ethi	Africa	231k	260	863
en_af	Afrikaans	Latn	Africa	219k	130	863
en_ber-Latn	Tamazight (Latin Script)	Latn	Africa	206k	130	862
en_ber	Tamazight (Tifinagh Script)	Tfng	Africa	206k	130	862
en_ee	Ewe	Latn	Africa	202k	130	863
en_pcm	Nigerian Pidgin	Latn	Africa	195k	130	864
en_yue	Cantonese	Hant	Asia	195k	584	863
en_kri	Krio	Latn	Africa	188k	130	863
en_tn	Tswana	Latn	Africa	182k	66	863
en_ve	Venda	Latn	Africa	167k	66	863
en_bm-Nkoo	NKo	Nkoo	Africa	167k	66	863
en_bem	Bemba (Zambia)	Latn	Africa	166k	66	863
en_ts	Tsonga	Latn	Africa	165k	66	863
en_tum	Tumbuka	Latn	Africa	164k	66	863
en_ss	Swati	Latn	Africa	163k	66	863
en_ktu	Kituba (DRC)	Latn	Africa	162k	66	863
en_nr	South Ndebele	Latn	Africa	159k	66	863
en_fon	Fon	Latn	Africa	157k	66	863
en_ndc-ZW	Ndau	Latn	Africa	156k	66	863
en_kg	Kongo	Latn	Africa	154k	66	863
en_dov	Dombe	Latn	Africa	153k	66	863
en_nd	North Ndebele	Latn	Africa	150k	66	863
en_ki	Kikuyu	Latn	Africa	149k	66	863
en_lu	Kiluba (Luba-Katanga)	Latn	Africa	148k	66	863
en_efi	Efik	Latn	Africa	147k	66	863
en_cgg	Kiga	Latn	Africa	147k	66	863
en_din	Dinka	Latn	Africa	145k	66	863
en_rn	Rundi	Latn	Africa	144k	66	863
en_tiv	Tiv	Latn	Africa	141k	66	863
en_kr	Kanuri	Latn	Africa	139k	66	863

Lang. pair	target language name	ISO 15924 Script	Continent	trg.chars	S.DOC	S.SENT
en_alz	Alur	Latn	Africa	139k	66	863
en_mfe	Mauritian Creole	Latn	Africa	137k	66	863
en_dyu	Dyula	Latn	Africa	136k	66	863
en_ach	Acholi	Latn	Africa	135k	66	863
en_dje	Zarma	Latn	Africa	135k	66	863
en_aa	Afar	Latn	Africa	133k	66	863
en_bci	Baoulé	Latn	Africa	131k	66	863
en_sus	Susu	Latn	Africa	128k	66	863
en_gaa	Ga	Latn	Africa	126k	66	863
en_mos	Mooré	Latn	Africa	125k	66	863
en_aeb	Tunisian Arabic	Arab	Africa	115k	66	862
en_lij	Ligurian	Latn	Europe	114k	25	863
en_apd	Sudanese Arabic	Arab	Africa	112k	66	855
en_ayl	Libyan Arabic	Arab	Africa	109k	66	863
en_scn	Sicilian	Latn	Europe	102k	100	0
sw_zh	Mandarin Chinese	Hans	Asia	101k	330	0
en_kl	Kalaallisut	Latn	Americas	97k	0	863
am_zh	Mandarin Chinese	Hans	Asia	96k	329	0
en_es	Spanish	Latn	Europe	88k	0	863
en_sat	Santali (Ol Chiki script)	Olck	Asia	83k	0	863
en_bo	Tibetan	Tibt	Asia	82k	0	863
en_lus	Mizo	Latn	Asia	82k	0	863
en_gn	Guarani	Latn	Americas	82k	0	863
en_ay	Aymara	Latn	Americas	82k	0	863
en_sat-Latn	Santali (Latin Script)	Latn	Asia	81k	0	863
en_hac	Hawrami	Arab	Asia	77k	0	863
en_glk	Gilaki	Arab	Asia	77k	0	863
en_ckb	Sorani	Arab	Asia	77k	0	863
en_is	Icelandic	Latn	Europe	77k	0	863
en_sa	Sanskrit	Deva	Asia	77k	0	863
en_qu	Quechua	Latn	Americas	74k	0	863
en_brx	Bodo (India)	Deva	Asia	74k	0	863
en_ks	Kashmiri	Arab	Asia	73k	0	863
en_pa-Arab	Lahnda Punjabi (Pakistan)	Arab	Asia	73k	0	863
en_mni-Mtei	Meiteilon (Manipuri)	Mtei	Asia	71k	0	863
en_ks-Deva	Kashmiri (Devanagari script)	Deva	Asia	65k	0	863

Table H.1: Details on all SMOL language pairs, sorted by the total number of characters in the target side (col. 5). The last two columns are the number of examples per language pair; keep in mind that an example for SMOLSENT is a sentence pair but for SMOLDOC is a document/paragraph. Language pairs are only listed in the direction in which they were translated, so no xx→en pairs are present.

I Data sample

I.1 Sample datum from SmolSent

```
{'id': 381,
 'sl': 'en',
 'tl': 'luo',
 'is_src_orig': True,
 'src': 'Rih, a deaf former soldier, plots rebellion while married to a queer,
    teenage god.',
 'trg': 'Rih, mane en jalweny ma Radin, ochano balo ka koni to okendo ng'ano manigi
    kido mar chuech kamare, nyasaye ma en ojana.'
}
```

I.2 Sample datum from SmolDoc

```
{
'id': 'topic_587__weyiwiniwaaotiwenwy',
 'sl': 'en',
 'tl': 'pcm',
 'is_src_orig': True,
 'factuality': 'ok', # this is a story so there is no factual claim that could be
    wrong
 'srcs': ['"What the hell are you doing, you idiot?!"',
       '"Excuse me?"',
       '"You cut me off! You almost made me crash!"',
       '"I'm sorry, I didn't mean to. I was just trying to get around that slow-
          moving truck."',
       '"Well, you could have at least used your turn signal!"',
       '"I did use my turn signal!"',
       '"No, you didn't! You just pulled right out in front of me!"',
       '"I'm telling you, I used my turn signal!"',
```

```
        '"Whatever. You're still a terrible driver."',
        '"And you're a jerk!"',
        '"At least I know how to drive!"',
        '"Oh, yeah? Well, I'm a better writer than you are!"',
        '"That's debatable."',
        '"It's not debatable! I'm Ernest Hemingway!"',
        '"Who?"',
        '"Ernest Hemingway! The greatest writer of all time!"',
        '"Never heard of him."',
        '"Well, you've heard of me now!"',
        '"Yeah, I heard of you."'],
 'trgs': ['"Wetin di hell dey do, yu idiot?!"',
        '"Ekskuse mi?"',
        '"Yu komot mi! Yu almost make mi krash!"',
        '"I dey sorry, I nor wont do am. I just dey try get around dat truk wey slow
            ."',
        '"Well, yu for don yus yor turn sign!"',
        '"I yus mai turn sign!"',
        '"No, yu nor turn am! Yu just turn rite in front of mi!"',
        '"I dey tell yu, I yus mai turn sign!"',
        '"Wateva. Yu still bi one tribol driva."',
        '"And yu bi jerk!"',
        '"At least I sabi hau to drive!"',
        '"Oh, yeah? Well, I bi ogbonge writa pass yu!"',
        '"Wi fit dibate dat."',
        '"nortin to dibate! I bi Ernest Hemingway!"',
        '"Who?"',
        '"Ernest Hemingway! De writa of all taim wey grate pass!"',
        '"Neva hear am."',
        '"Well, yu don hear mi nau!"',
        '"Na so, I don hear yu."']
}
```

Improved Norwegian Bokmål Translations for FLORES

Petter Mæhlum
Language Technology Group
University of Oslo, Norway
pettemae@ifi.uio.no

Anders Næss Evensen
Disputas
anders@disputas.no

Yves Scherrer
Language Technology Group
University of Oslo, Norway
yves.scherrer@ifi.uio.no

Abstract

FLORES+ is a collection of parallel datasets obtained by translation from originally English source texts. FLORES+ contains Norwegian translations for the two official written variants of Norwegian: Norwegian Bokmål and Norwegian Nynorsk. However, the earliest Bokmål version contained non-native-like mistakes, and even after a later revision, the dataset contained grammatical and lexical errors. This paper aims at correcting unambiguous mistakes, and thus creating a new version of the Bokmål dataset. At the same time, we provide a translation into Radical Bokmål, a sub-variety of Norwegian which is closer to Nynorsk in some aspects, while still being within the official norms for Bokmål. We discuss existing errors and differences in the various translations and the corrections that we provide.

1 Introduction

This paper describes our submission to the WMT 25 open language data shared task, where participants were asked to contribute to open dataset collections such as FLORES+, the MT Seed dataset or other parallel datasets. We have chosen to focus on the Norwegian Bokmål part of the FLORES+ dataset, as the authors notice non-fluencies in the dataset in 2024, and notified the original authors. These issues were attempted resolved in a process that lead to additional errors, which are the ones that form the basis of this paper. In addition to correcting these translations, we translate the resulting Norwegian Bokmål dataset into a version of a specific variety of written Norwegian called radical Bokmål. Having these two normed varieties can be beneficial for experiments where variation in Norwegian spelling norms is important. We summarize some of the encountered errors in the newest Bokmål translations, and show some results on several machine translations baselines for the new and existing Norwegian versions.

2 The Norwegian Language and its Writing Norms

Norwegian is one of the official languages of Norway, along with Sámi languages and Norwegian Sign Language. It is a North-Germanic language historically descendant of Western Norse, but following large Saxon and East Norse influences, is largely mutually intelligible with its neighbors Swedish and Danish, and more different from Icelandic and Faroese. However, following centuries of having Danish as Norways national language, nationalist movements in the late 19^{th} century lead to the establishment of two written standards: Landsmål (today **Nynorsk**), which was based on dialects "untainted" by Danish, and Rigsmål (Riksmål, today **Bokmål**), which was Norwegianized Danish. Nynorsk historically aimed at preserving Norwegian-specific features, which means that Saxon and Danish influences are less pronounced in Nynorsk than in Bokmål.

2.1 Conservative, Moderate and Radical Bokmål

Within both written norms however, considerable variation exists. This variation is not arbitrary, and generally follows typical patterns, leading to what is known as **norm clusters** (nor. *normklynger*) (Dyvik, 2009). On the Bokmål side, perhaps the most common sub-norm is **Moderate Bokmål** (MBM) (or Conservative Bokmål, CBM)[1], which is what dominates especially formal Norwegian discourse. This variety is known for Danish-like conjugational and declensional patterns: preterite in *-et*, no feminine nominal endings (with a few exceptions). On the other hand, **Radical Bokmål** (RBM) aims at a style closer to how many people in

[1]While the terms *moderate* and *conservative* are used synonymously by some authors, some prefer to reserve *conservative* for the most extreme (Danish-like) Norwegian, and reserve *moderate* for a norm that allows for some radical elements.

Proceedings of the Tenth Conference on Machine Translation, pages 1124–1132
November 8-9, 2025

(Eastern) Norway speak, and adopts features shared with Nynorsk (NN) in some regards: obligatory feminine marking for articles and noun declension, preterite in *-a* (NN also *-a*) where MBM has *-et*, and neuter definite plurals in *-a* (NN also *-a*) where MBM has *-ene*. There are also some sound correspondences, with a difference between diphthongs and monophthongs being especially common. For example, RBM has *mjølk* where MBM has *melk* (NN *mjølk)*. As illustrated in the artificial example below, this affects both morphology and syntax, with Nynorsk added for comparison. Note how while RBM is said to be closer to Nynorsk, this is not to say that all RBM forms exist in Nynorsk, as exemplified by *skog* (MBM, NN) and *skau* (RBM).

(1) MBM: I helgen var jeg på hytten i skogen og danset med min søster.

RBM: I helga var jeg på hytta i skauen og dansa med søstera mi.

NN: I helga var eg på hytta i skogen og dansa med søstera (or systera) mi.

The degree to which a writer follows these patterns differs, leading to many possible variations within this spectrum. In our efforts to provide a radical form, we chose the most radical form as presented in the official Norwegian dictionary *Bokmålsordboka*.[2]

3 The FLORES Dataset

The FLORES dataset is an evaluation dataset for multilingual machine translation, consisting of a *dev* and a *devtest* part with about 1000 sentences each. The dataset is multiparallel and English-centric: the original sentences are in English, and all other language variants were produced by translation. Several versions were made available over time, reflecting efforts to increase language coverage and address quality issues.

FLORES101 was the first version of FLORES, covering 101 languages, including Norwegian (Goyal et al., 2021). While the authors claimed that the sentences were "[...] translated in 101 languages by professional translators through a carefully controlled process", we observed severe quality problems with the Norwegian Bokmål translations. See further discussions in 3.1.

[2]Bokmålsordboka. The Language Council of Norway (Språkrådet) and the University of Bergen https://ordbokene.no.

FLORES200 was a continuation from both FLORES101 and Guzmán et al. (2019), with an increased coverage of 200 languages (NLLB Team, 2022). The Norwegian sentences appear to be unchanged between FLORES101 and FLORES200. The FLORES200 translations were used, among others, in the Belebele (Bandarkar et al., 2024) benchmark. Quality problems in the former therefore directly affect results reported on the latter dataset. We have used the Bokmål sentences from FLORES200 both as an aid in correcting the translations, and as a point of comparison agains the new dataset as a whole. These sentences initially struck the authors as unnatural, with exampled reported such as translating *iron* (the metal) as *strykejern* (eng. 'clothes iron'), and (judicial) *court* as *hoff* (eng. royal court).

FLORES+ The responsibility for the FLORES datasets was eventually moved to the Open Language Dataset Initiative.[3] As a result, the updated versions are referred to by FLORES+ and published on HuggingFace.[4]

In January 2024, the authors of this paper reached out to the original FLORES101 authors to express concern over the quality of the Norwegian Bokmål dataset, based on FLORES200. Following this, the dataset was updated, as indicated by a changelog note from November 11^{th} 2024[5]. This note informs that the Norwegian version has been updated after quality assessment, but with no further information. Going through these changes, we see, however, that not all errors were corrected, and that new ones were introduced. Correcting these errors is the main focus of this paper.

3.1 Problems with the Norwegian FLORES Translations

The Norwegian version published as part of FLORES101 (referred to as **BM1**) contained a range of issues, some of which were amended in the FLORES+ version (referred to as **BM2**).

When comparing the BM1 and BM2, the naturalness of the sentences have improved in some cases, but there are still multiple mistakes left in the dataset. Overall, BM1 used a syntax that was less influenced by the original English. In some

[3]https://oldi.org/

[4]https://huggingface.co/datasets/facebook/flores

[5]https://huggingface.co/datasets/openlanguagedata/flores_plus/blob/main/CHANGELOG.md#20--2024-11-11

cases, BM2 is even worse than BM1. Every single sentence in the Bokmål dataset was changed during this edit, but it seems like the BM2 translations show signs of not having referenced the BM1 translations, as in several cases where both Nynorsk and BM1 have a correct translation, but BM2 still mistranslates, for example the English *engraver*, which is translated correctly as *gravør* in NN and BM1, is erroneously translated as *graver* 'digger' in BM2. While the Nynorsk dataset also had changes in 21 sentences, this turned out to only be differences in trailing spaces.

Certain errors are so pronounced that it is difficult to conceive they were produced by a professional translator, or even by a fluent speaker of Norwegian. This can be illustrated by example 2 where the English word 'bill' has been translated as *lovforslag* 'bill (judicial)', while in the new translation it has been changed to *regning* 'bill, receipt'. The terms are not ambiguous as in English, and the result is comical.

(2) BM1: Det opprinnelige **lovforslaget** ble utarbeidet av tidligere ordfører i São Paulo [...]

BM2: Den opprinnelige **regningen** ble utarbeidet av tidligere borgermester i São Paulo [...]

EN: The original **bill** was drafted by former mayor of São Paulo, Marta Suplicy.

These examples make it clear that despite the corrections being made in 2024, not all mistakes were fixed, and some new ones were introduced. This prompted us to critically assess the BM2 translations and correct them. In the following section we describe our correction process, then follow with a brief overview of error types and key statistics. We refer to our corrected Bokmål translations of FLORES+ as **BM3**, and to the Radical Bokmål version as **BM3R**.

4 Translation Correction

We introduce our methodology and discuss some of the encountered errors. See Appendix A for selected example sentences in all languages involved in this process.

4.1 Methodology

Our aim for this effort is not a full retranslation of the English, but rather to take the BM2 translations as a starting point. We assume that the BM2 translations follow the FLORES translation guidelines[6], which do not allow any AI tools. They should therefore provide an appropriate point of departure, being the most recent translations. We aim mostly at correcting *obvious* grammatical and lexical mistakes in the dataset, while keeping the structure and otherwise correct lexical choices of the original translators intact. However, in more severe cases, the other sentences were used as reference, especially the NN sentences, as they overall hold a much higher quality level, though not free of errors. Some concrete breaches of the translation guidelines were also corrected, notably cases where units of measure were translated and converted. There were also cases of named entities that were not changed, even though an established Norwegian spelling exists, such as Eng. *Pythagoras* vs. Nor. *Pytagoras*.

Measured in error correction on the most recent Bokmål translations, 64.4% of the sentences in the devtest split contained errors that were fixed, with 70.8% for the dev split. We make an attempt at avoiding corrections due to matters of choice, but do correct in the following cases:

1. Grammatical mistakes
2. Clearly mistranslated terms
3. Orthographic or punctuation mistakes
4. Misunderstanding of context, etc.
5. Mismatching radicalness

In some cases, errors in the Nynorsk were discovered, but they were not corrected. We urge others to revisit the Nynorsk translations.

Two annotators worked on the correction task, with about 25 hours of work for each annotator. Both were native Norwegian speakers with some background in professional translation. The datasets were split in two, with each person correcting their half, before quality checking the other person's corrections. It was not the translators' intention to re-translate, simply to correct the existing translations, basing the new translations on these existing ones to provide a more fluent and correct sentence, and thus improving the overall quality of the dataset. Changes in radicalness were only done in cases where it did not match the dataset overall.

In the rare cases where none of the earlier translations are correct, the translators allow themselves

[6] https://oldi.org/translation-guidelines.pdf

to retranslate. This is mostly seen in the case of specific jargon or common misunderstandings, in places where the English syntax is kept in the translations despite being ungrammatical, or if there were no appropriate translations for them to base themselves on. Therefore some cases with English-like syntax but no grammatical or lexical mistakes have been kept.

When the communicative intent of the English sentence is not hindered by small errors, these errors are not carried over in the Norwegian translation.

Although difficult to avoid completely, the translators have attempted to avoid letting stylistic preferences affect the correction. For example in cases where the translations is passable, but the translator would prefer another word, we have avoided correcting them, except in cases where the sentence sounds very unnatural. All cases were discussed between the two annotators. In some borderline cases, where it is difficult to argue in disfavor of the original translation, the sentences are kept, as with the cases discussed above. Semi-authoritative sources such as the Norwegian Wikipedia, the Norwegian encyclopedia Store Norske Leksikon (SNL) or books in the National Library (NB) collection were used to guide term usage.[7]

4.2 Types of errors

In order to give an impression of some of the mistakes found necessary to correct by the annotators, we attempt to summarize some of the more common ones in broad groups.

Term Coinage and Anglicisms The original translator has in several instances made up words with little to no previous usage, in cases where there are clearly preferred terms. Examples include *meteordusj* 'meteor shower' for meteorsverm lit. 'meteor swarm', or in the case of *martianer* in 3, which needs to be rewritten as *fra Mars*.

(3) [...] rundt 34 var **martianer** i opprinnelsen.

[...] about 34 have been verified to be **martian** in origin.

Direct Translation Similar to coinages, but where a coinage might be seen as a creative way to translate a term into Norwegian, where the translator is unaware of an existing or more commonly used term, the direct translations (ie. translating word-by-word) lead to clear errors, as the resulting word has a completely different meaning in Norwegian. Sometimes these translations might pass as understandable anglicisms, while at times they are nonsensical. This is especially typical for fixed expressions. In 4, the English phrase *on its own* is translated as *alene* 'alone', which does not share the same use. In 5, the English *in a big way* is translated to Norwegian *på en stor måte*, which does not make sense.

(4) Madagaskar er den klart største, og et kontinent **alene** når det gjelder dyreliv.

Madagascar is by far the biggest, and a continent **on its own** when it comes to wildlife.

(5) Araberne førte også islam til landene, og det tok **på en stor måte** i Komorene og Mayotte.

The Arabs also brought Islam to the lands, and it took i**n a big way** in the Comoros and Mayotte.

Misunderstanding Context In a few cases, a Norwegian word might be a possible translation of an English word, but the translator misunderstands the context, and translates the wrong sense of the word. This is different from the two above in that the word is actually correct, but not the correct translation in this case and context. In 6, the English *peers* is translated as NB *jevnaldrende* (lit. same-aged), a term used especially when discussing peers in primary school and similar cases. It cannot be used in the sense of professional peers, which is the intended meaning here. In 7, the understood context of the elided 'flair' has caused the translator to use the word *afrikaner*, 'African (person)', instead of the adjective. Most lexical errors fall in this category.

(6) Generelt sett kan to atferd oppstå når ledere begynner å lede sine tidligere **jevnaldrende**.

Generally speaking, two behaviors can emerge as managers begin to lead their former **peers**.

(7) [...] fordi den har mer arabisk stil enn en **afrikaner**.

[...]because it has more of an Arabic flair than of an **African**.

[7] Wikipedia:https://no.wikipedia.org/wiki/Forside, SNL: https://snl.no/, NB: https://www.nb.no/

Agreement Norwegian Bokmål exhibits agreement between some parts of speech where English does not, such as past participles, adjectives and some determiners. The authors found cases of mismatching agreement in the BM2 texts, such as "en naturlig forekommende encellede marine organisme.", where *encellede* 'single cell' and *marine* 'marine' are plural forms, while *en* 'a' and *organisme* 'organism' are singular. The correct form would be *encellet* and *marin*.

Subject-Possessor Mismatch These are cases when the translator fails to use the correct possessor. This is especially clear in Norwegian, as there is a difference in whether the grammatical subject of a sentence is the possessor or not. In 8 the third person possessor *sitt* is ungrammatical due to "nylige eksempler på [...] arbeid" being the subject of the subordinate clause, and in this case, the possessor should have been *hans* (eng. his).

(8) **Han** var også engasjert i gravering av sedler i mange land, nylige eksempler på **sitt** arbeid, inkludert statsministerportretter [...]

He was also engaged in engraving banknotes for many countries, recent examples of **his** work including the Prime Ministerial portraits [...]

Incorrect Noun Gender Several nouns have been used with the wrong grammatical gender, for example, *sexet* (neut.) 'the sex', instead of *sexen* (masc.), and *giftet* (neut.) 'the poison' instead of *giften/gifta* (masc./fem.) and *det største anskaffelsen* (neut.) 'the largest acquisition', instead of *den store anskaffelsen* (masc.). In the latter case, the mismatch is especially clear, as the noun is already declined in the masculine definite form, causing an agreement error.

The gender mismatch also extends to anaphoric pronouns, as in ***Et motbåtskip*** *av Avenger-klasse [...].* ***Den*** *er tildelt [...]*, where the first noun is neuter, while the referring pronoun is common gender. This is different from the agreement point above, in that these cases mistake the gender of the nouns themselves, not just the modifying elements.

Syntactical errors While English is syntactically close to Norwegian, there are certain constructions that when directly translated become ungrammatical or unnatural. An example is seen in 9, where in English an appositioned place name can function as an indicator of origin, while this has to be rewritten in Norwegian, for example as *jazz-spiller fra Utah* 'Jazz player from Utah', or in 10, where the subordinate clause initialized by a single past participle is very marked in Norwegian and must usually be rewritten.

(9) NBAs beslutning fulgte en **Utah Jazz-spiller** [...]

The NBA's decision followed a **Utah Jazz player** testing positive for the COVID-19 virus.

(10) **Født i Hong Kong** studerte Ma ved New York University [...]

Born in Hong Kong, Ma studied at New York University

Repetitions A final type of error is the case where multiple words in English have been translated to the same word in Norwegian. An example is seen in 11. These need to be rewritten to avoid repetition if no good synonyms can be found in Norwegian.

(11) ulovlige handlinger som dommere, **advokater, advokater og advokater** har gjort i løpet av de foregående årene.

[...] illegal actions that judges, **lawyers, solicitors and attorneys** have done during the previous years.

Minor Mistakes In addition to the issues above, there are other minor mistakes. However, one striking aspect about all of these is that they are rarely encountered with native speakers, as these are core components of Norwegian grammar, and not infrequent or rare phenomena.

4.3 Other Correction Issues

When correcting, we have focused mainly on improving the latest Bokmål translations, as these are supposed to be improvements on the earlier translations. They are mostly in a standard moderate variety, allowing for feminine definite forms of very frequent feminine nouns, as is typical in moderate BM, but varying in consistency when it comes to some less frequent words and other potentially radical forms.

In the case of many loanwords in Norwegian, both older spellings and more Norwegianized spellings are allowed in some cases. For many old loanwords, these spellings are naturalized and

one might even think about their original spellings, such as *byrå* (fr. bureau) 'office' and *sjåfør* (fr. chauffeur) 'driver', while for many more recent words, especially from English, the Norwegian alternatives can sometimes be more marked. The translators of the most recent Bokmål dataset seem to have a preference for keeping original spellings, and we have kept them. However, in the radical versions, we have used the more Norwegian versions, leading to differences such as streame vs. strømme 'to stream', container vs. konteiner 'container', etc.

4.4 Radical version

Following the correction of the BM2 version into BM3, we then convert these into radical Bokmål (BM3R).

As described above, Radical Bokmål is a subvariety of Bokmål, where radical options are chosen. These options refer to lemma varieties sanctioned by the dictionary Bokmålsordboka, while trying to stay as close as possible to the normative guidelines put forth by the association for Radical Bokmål[8]. The differences broadly fall into three categories:

Sound Correspondences Many optional forms are based on differences in diverging sound changes that are semi-regularly executed in Bokmål. These are found broadly across parts of speech, as in *melk* (MBM)/*mjølk* (RBM) and *fløte* (MBM) and *fløyte (RBM-*[9]

(12) [...] masse krem**fløte** (ikke **melke**skum) og te blir servert uten **melk**. (MBM)

[...] masse krem**fløyte** (ikke **mjølke**skum) og te blir servert uten **mjølk**. (RBM)

Another large category is morphology. In our case, this especially applies to the endings discussed above, in addition to using the deverbal nominal suffix *-ing* instead of the more MBM-coded *-else*, where these are listed as equal in the Bokmål dictionary.

(13) [...] en viktig del av **opplevelsen**. (MBM)

[...] en viktig del av **opplevinga**. (RBM)

Finally, the only syntactic change is when rewriting possessive constructions with a bare genitive 's' in MBM, which are preferred as prepositional phrases in RBM, as in 14.

(14) **Tigerens brøl** er ikke [...] (MBM)

Brølet til tigeren er ikke [...] (RBM)

[8] https://bokmal.no/

[9] Note that marking a word as MBM/RBM in this case is for convenience, but it is not the case that all radical versions are equally strong indicators of RBM as this is a continuum.

BLEU	BM1	BM2	BM3	BM3R
BM1	–	32.31	35.59	32.54
BM2	32.37	–	73.73	64.89
BM3	35.63	73.65	–	86.89
BM3R	32.56	64.79	86.84	–
chrF	**BM1**	**BM2**	**BM3**	**BM3R**
BM1	–	62.65	63.76	62.22
BM2	61.58	–	84.81	81.51
BM3	62.96	85.21	–	94.77
BM3R	61.32	81.74	94.59	–

Table 1: Similarity metrics for the various translations using BLEU and chrF (rows refer to hypotheses, columns to references).

	MADLAD		NLLB		OPUS-MT	
Ref.	BLEU	chrF	BLEU	chrF	BLEU	chrF
BM1	33.44	62.03	31.31	59.53	34.91	62.96
BM2	62.11	79.72	58.98	76.82	69.50	84.38
BM3	56.12	75.59	52.01	72.16	60.06	77.75
BM3R	49.83	72.88	46.12	69.62	53.01	74.81

Table 2: Translation scores for three Bokmål system outputs, using the four different reference translations.

5 Experiments

5.1 Distances between translations

To quantify the differences between the Bokmål translations, we compute BLEU and chrF scores between pairs, taking one as a reference and the other as the hypothesis. Table 1 presents the results.

We observe that BM1 is most different from all other translations, with differences of around 30 BLEU points and 20 chrF points. BM3 is relatively similar to BM2, which is not surprising due to the majority of corrections being done on these sentences. The highest similarities are observed between the moderate and radical BM3 translations, suggesting that lexical and morphosyntactic contexts that allow variation are relatively rare in the dataset. Both metrics follow roughly the same pattern.

5.2 Machine translation experiments

We measure the impact of the updated translations on machine translation evaluation in two experiments, using English-to-Bokmål and Bokmål-to-English MT systems, respectively.

First, we translate the English FLORES dev set to Norwegian Bokmål and evaluate the output on all four reference translations. This shows how much the evaluation scores of a single translation can vary according to the reference used. We use three MT systems to translate from English to Bokmål: MADLAD-400-3B-MT[10], NLLB-200-distilled-1.3B[11], and OPUS-MT[12]. Table 2 shows the results.

It can be seen that BM1 provides the lowest translation scores, suggesting that the reference translations are too different from the ones produced by off-the-shelf MT systems. On the other hand, the highest scores are obtained when using BM2 as a reference; this could hint to increased translationese in this dataset.

The moderate reference yields higher scores than the radical reference, which suggests that the translations produced by the three MT systems is more similar to the (more widely used and less marked) moderate variant. The radical reference impacts BLEU score slightly more than chrF score, as many moderate/radical differences occur at subword level (e.g., inflectional endings).

The three MT models provide output of similar quality, with OPUS-MT outperforming MADLAD and NLLB. Interestingly, the score differences across models are more pronounced with BM2 and BM3 than with BM1. BM1 seems therefore of limited usefulness to discriminate between different MT systems.

Second, we use the four Bokmål translations as input to Norwegian-to-English translation systems and evaluate the English outputs using the English FLORES reference. We again use MADLAD-400-3B-MT and NLLB-200-distilled-1.3B, as well as an OPUS-MT model covering the opposite translation direction[13]. The results are presented in Table 3.

[10] https://huggingface.co/google/madlad400-3b-mt

[11] https://huggingface.co/facebook/nllb-200-distilled-1.3B

[12] https://huggingface.co/Helsinki-NLP/opus-mt-tc-bible-big-deu_eng_fra_por_spa-gmq

[13] https://huggingface.co/Helsinki-NLP/opus-mt-tc-big-gmq-en

	madlad3b		nllb1.3bdist		opusmt	
Input	BLEU	chrF	BLEU	chrF	BLEU	chrF
BM1	43.69	68.02	40.48	64.73	44.10	68.26
BM2	63.98	80.68	58.97	76.82	66.66	81.94
BM3	56.18	75.24	52.19	71.87	58.38	76.44
BM3R	55.16	74.73	51.08	71.13	57.61	75.85

Table 3: Translation scores for English system outputs produced from various Bokmål inputs.

The results quite closely reflect those of the English-to-Norwegian translation, with the exception of the moderate/radical distinction, which almost has no impact on the models' capacity to translate the text to English.

We find that the initial translation (BM1) severely underestimates the models' true MT capabilities, both when used as a reference and as an input text. The updated version provided by the FLORES team (BM2) yields the highest scores, whereas the translations provided in this paper (BM3) lie in between.

An explanation for some of this effect, is that we still observe translationese tendencies in the translated texts, for example, all three systems provide the following translation in example 15 (madlad3b adds *av*, the others do not). The word (chemical) *element* is *grunnstoff* in Norwegian, and anglicisms like this might inflate the scores for the earlier translations.

(15) Du kan også ha legeringer som inneholder små mengder (av) ikke-metalliske elementer som karbon. (PRED)

You can also have alloys that include small amounts of non-metallic elements like carbon. (EN)

Det finnes også legeringer som inneholder små mengder ikke-metalliske grunnstoffer som karbon. (GOLD-MBM)[14]

6 Conclusion

Even after its initial correction, several obvious and non-native-like mistakes remained in the FLORES+ Bokmål dataset. Our attempt has corrected the most obvious mistakes, making sure that there are at least no grammatical or lexical mistakes in the dataset, without introducing excessive changes to

[14] Adding *av* is acceptable.

the work done by the professional translators. We hope that these corrections make results from these datasets more reliable.

On a more personal note, this is not the first time the authors experience problems with context and understanding coming in the way when translating datasets that are supposed to be the basis of massive-parallel datasets. We urge the creators of such original datasets to perhaps add clarifying remarks where there might be misunderstandings. Following the observation that close to 70% of all sentences in the corrected dataset contained at least one lexical or grammatical error, we recommend earlier users of the dataset to reevaluate results used on this dataset. There is also some reason to doubt the claims that all these translations were indeed done by professional translators, and we hope that future dataset creators will use the native professional communities to gain valuable feedback in these situations.

Limitations

We observe that the Nynorsk overall holds a much better quality, but that this dataset would also benefit from a round of corrections by someone qualified. We urge a native Nynorsk writer with translation experience to do a similar check of the Nynorsk data. We also acknowledge that some cases were difficult to translate due to a lack of domain knowledge, especially alongside ambiguous English original sentences, and that the focus of this effort was to remove clear errors.

Acknowledgments

This work was supported by the HPLT project which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350. The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

References

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Helge Dyvik. 2009. Å navigere i skriftspråkets rom. om normklynger i bokmål og nynorsk. *Språknytt*, 37(3):15–21.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. Two new evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

A Dataset example

Selected examples from the corrected datasets, along with BM1 (FLORES200), BM2 (FLORES+), corrected version (BM3), radical version (BM3R) and Nynorsk (NN).

EN	BM1	BM2	BM3	BM3_RAD	NN
The truck driver, who is aged 64, was not injured in the crash.	Lastebilsjåføren på 64 år, ble ikke skadet i styrten.	Lastebilsjåføren, som er 64 år gammel, ble ikke skadet i ulykken.	Lastebilsjåføren, som er 64 år gammel, ble ikke skadet i ulykken.	Lastebilsjåføren, som er 64 år gammel, ble ikke skada i ulykka.	Lastebilsjåføren på 64 år vart ikkje skada i kollisjonen.
During his trip, Iwasaki ran into trouble on many occasions.	I løpet av reisen sin kom Iwasaki i trøbbel ved flere begivenheter.	Under reisen hans kom Iwasaki i vanskeligheter ved mange anledninger.	Under reisen sin kom Iwasaki i vanskeligheter ved mange anledninger.	Under reisa sin kom Iwasaki i vanskeligheter ved mange anledninger.	Under turen hans møtte Iwasaki på problem ved fleire høve.
In just two weeks the Americans and Free French forces had liberated southern France and were turning towards Germany.	I løpet av bare to uker hadde amerikanerne og frie franske styrker frigjort den sørlige delen av Frankrike og vendte seg mot Tyskland.	På bare to uker hadde amerikanerne og de franske franske styrkene frigjort Sør-Frankrike og vendte seg mot Tyskland.	På bare to uker hadde amerikanerne og de frie franske styrkene frigjort Sør-Frankrike og vendt seg mot Tyskland.	På bare to uker hadde amerikanerne og de frie franske styrkene frigjort Sør-Frankrike og vendt seg mot Tyskland.	På berre to veker hadde amerikanarane og sjølvstendige franske styrkar frigjort den sørlege delen av Frankrike, og var på veg mot Tyskland.

NRC Systems for the WMT2025-LRSL Shared Task

Samuel Larkin **Chi-kiu Lo 羅致翹** **Rebecca Knowles**
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
{samuel.larkin,chikiu.lo,rebecca.knowles}@nrc-cnrc.gc.ca

Abstract

We describe the NRC team systems for the WMT25 Shared Tasks on Large Language Models (LLMs) with Limited Resources for Slavic Languages. We participate in the Lower Sorbian and Upper Sorbian Machine Translation and Question Answering tasks. On the machine translation tasks, our primary focus, our systems rank first according to the automatic MT evaluation metric (chrF). Our systems underperform on the QA tasks.

1 Introduction

This paper describes our systems submitted to the WMT 2025 LLMs with Limited Resources for Slavic Languages Shared Task. We focused primarily on LLM-based machine translation (MT) into Upper Sorbian (hsb) and Lower Sorbian (dsb). Balancing two competing tasks typically comes at the cost of performance on one task or the other. While we performed preliminary experiments on QA, our initial results were unsatisfactory and we chose to submit a system that was trained only for MT as our primary submission.

2 Data and Models

We constrained our systems to the corpora offered by the organizers. The WMT25 corpora are available in the shared task github repository[1] while the previous years' corpora are available in the WMT22 repository.[2] MT data is described in more detail in Section 3.1 while QA data is described in more detail in Section 4.1.

As required by the shared task, we trained our models using the 0.5B, 1.5B, and 3B size Qwen2.5 models as the base. Our submitted system for the shared task is based on `Qwen2.5-1.5B-Instruct`.[3]

3 Machine Translation

We focused primarily on MT performance in our submissions. In this section, we describe preliminary experiments on MT, leading to the choices we made for our final submissions (Section 5). We used the `doc_to_text` task prompt suggested in the `lm-eval` test harness:[4]

```
Translate the following German
text to Lower Sorbian.  Put it in
this format <dsb> Lower Sorbian
translation </dsb>.\n<deu> {{de}}
</deu>
```

as our machine translation prompt for Lower Sorbian (replacing `Lower Sorbian` by `Upper Sorbian` and `dsb` by `hsb` in the Upper Sorbian setting). In addition to exploring hyperparameters for supervised finetuning using `LLaMa-Factory`, we tested which combinations of language pairs and directions to use for training as well as different stopping criteria and prompt templates.

3.1 Data

We used almost all corpora for training except the *dev* from 2025 (see Table 1 for details).[5] We used *2025_dev* to evaluate translation quality. To avoid data contamination in our experimental evaluations

[1] https://github.com/TUM-NLP/llms-limited-resources2025

[2] https://github.com/mariondimarco/WMT22_UnsupVeryLowResMT_Data

[3] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

[4] https://github.com/TUM-NLP/wmt25-lrsl-evaluation/blob/main/lm_eval/tasks/wmt25-lrsl/sorbian/deu-dsb/deu-dsb.yaml

[5] *2022.train_dsb_hsb_62564.dsb-hsb*, which contains 62,565 sentence pairs with 673,781 Lower Sorbian words and 654,445 Upper Sorbian words was unintentionally omitted from training. This corpus is equivalent in size (in lines) to roughly 7% of the raw data we used for training, roughly 26% the size (in words) of the Lower Sorbian data we used for training, and roughly 8% the size (in words) of the Upper Sorbian data we used in training. Due to time constraints, we were unable to perform experiments to determine the extent to which including this might have changed model performance.

Proceedings of the Tenth Conference on Machine Translation, pages 1133–1142
November 8-9, 2025

on the *dev* data, we remove the 2025 *dev* set from all the training material using sentence pairs as the duplication key. Table 1 shows the number of sentences and words for each corpora before and after filtering. We sampled 200 sentences from the 2025 *dev* sets to produce *2025.dev_sample.de-dsb* and *2025.dev_sample.de-dsb* respectively. The smaller samples were used for evaluation during training to track training progress (without the cost of decoding the full *dev* set), while the full *dev* sets were used for our internal evaluations and model choices.

We preprocessed all MT corpora by removing control characters, removing carriage return (`\x0D`). We also collapsed tabs to space, folded multiple spaces into a single space and removed trailing spaces. See Appendix A for details.

3.2 Training

Training for MT was performed using `LLaMa-Factory`'s implementation of supervised finetuning (called the `sft stage`). We took inspiration from GemmaX2 (Cui et al., 2025) for our base hyperparameters. They provide their configuration in their github repository.[6] Table 2 lists the hyperparameters that we explored during training for MT; we did not perform a full grid search of all parameters. The specific values used in our final submission are given in Section 5. All experiments were performed on Tesla V100-SXM2-32GB.

3.3 Language Pairs and Translation Direction

In low-resource machine translation research, it has often been considered beneficial to train on as much language data as possible, sometimes incorporating training data from related languages. This has been a component of past shared tasks on Sorbian languages (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022). We wanted to explore this in the LLM setting: should we finetune LLMs separately for each target language's corpora or combine Upper Sorbian and Lower Sorbian corpora to train a single LLM that could translate into both languages?[7] We were also interested in determining how unidirectional training (training only to translate into Upper Sorbian and/or Lower Sorbian) compared to bidirectional training (training to translate into Upper Sorbian and/or Lower Sorbian as well as into German). Table 3 and Table 4 show that according to chrF,[8] the best-scoring combination of corpora and direction was a single system that incorporates both language pairs in both directions.[9] However, the chrF difference between the unidirectional training and the bidirectional training was consistently small or non-existent. Given the computational cost of training bidirectionally (twice as much data), we opted to use a unidirectional setup.

While running early experiments, we briefly considered LoRA (Hu et al., 2021) training, but observed substantially lower performance in the range of 12-15 BLEU and 7-10 chrF. For this reason, we continued to train full model weights in the remainder of our experiments.

3.4 Stopping Criterion

To improve training time, we can take advantage of the early stopping criterion which halts training if a given criterion or metric does not improve for a given number of gradient updates known as steps. Our baseline consists of training for exactly 5 epochs (`#epochs`), which we compared with a cross-entropy loss (`eval_loss`) approach where the model stops when the evaluation loss does not improve for 10 steps, and a `chrF` approach that stops when chrF has not improved for 10 steps. The models for these experiments were unidirectional (into Sorbian) translation models trained on the German to Upper or Lower Sorbian data as well as the Upper Sorbian–Lower Sorbian data.

Table 5 and Table 6 show that training for a fixed number of epochs always outperformed the other stopping criteria, yielding higher chrF scores. We used a fixed number of epochs for our final submitted system.

3.5 LLM Template

`LLaMa-Factory` offers multiple templates to format the input examples and prompt. We tested three of these built-in templates for the translation task: `qwen` (since the model we train is Qwen-

[6] https://github.com/xiaomi-research/gemmax/blob/main/scripts/sft.sh

[7] We included Lower Sorbian–Upper Sorbian parallel data in these experiments in addition to the German corpora: *2022.dev_dsb_hsb_new.dsb-hsb* and *2022.valid_dsb_hsb.dsb-hsb* are Lower Sorbian–Upper Sorbian corpora and because of that, during training, they were seen in both directions—effectively doubling their sentence count contribution.

[8] `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1`

[9] For completeness, we show Upper Sorbian systems' translations of Lower Sorbian text as well as the reverse, but show them in parentheses.

	raw				cleaned			
corpora	sentence	de_word	dsb_word	hsb_word	sentence	de_word	dsb_word	hsb_word
2020.devel.hsb-de.de-hsb	2000	24413		21692	1986	24246		21540
2020.devel_test.hsb-de.de-hsb	2000	24482		22081	1981	24241		21865
2020.train.hsb-de.de-hsb	60000	724572		639740	59703	720816		636414
2021.devel.dsb-de.de-dsb	601	7722	7231		601	7722	7231	
2021.devel_test.dsb-de.de-dsb	602	7786	7239		602	7786	7239	
2021.train.hsb-de.de-hsb	87521	1251339		1094421	86719	1240939		1085235
2022.40194_train_dsb_de.de-dsb	40194	514843	468509		40194	514843	468511	
2022.dev_dsb_hsb_new.dsb-hsb	700		7651	7416	700	700		
2022.HSB-DE_dev.tsv.de-hsb	2000	25607		22731	34	441		405
2022.HSB-DE_train.tsv.de-hsb	301536	3986351		3501610	301536	3986351		3501610
2022.valid.de-dsb	1353	9852	8424		1	2	3	
2022.valid_dsb_hsb.dsb-hsb	709		4592	4539	709	709		
2025.train.de-dsb	171964	2209255	2044017		171964	2209256	2044023	
2025.train.de-hsb	187270	2965088		2676309	187270	2965088		2676309
total	858450	11751310	2547663	7990539	854000	11703140	2527007	7943378
2025.dev.de-dsb	4000	46577	42295		4000	46577	42295	
2025.dev.de-hsb	4000	57580		51605	4000	57580		51605
2025.dev_sample.de-dsb					200	2291	2102	
2025.dev_sample.de-hsb					200	2855		2575

Table 1: Sentence count, per language word count of corpora. Training corpora are in the top section whereas the development sets are in the lower part of the table.

hyperparameter	values
per_device_train_batch_size	`4`, **`8`**, `32`, `64`
learning_rate	`1.0e-05`, `2.0e-05`, **`7.0e-05`**
num_train_epochs	`1`, `5`, `10`, **`20`**, `100`
gradient_accumulation_steps	`4`, **`8`**, `16`, `32`
max_grad_norm	**`1.0`**
warmup_ratio	**`0.0`**, `0.01`
weight_decay	**`0.0`**, `0.01`
lr_scheduler_type	**`cosine`**, `inverse_sqrt`
template	**`chatml`**, `empty`, `qwen`

Table 2: The values of hyperparameters that were investigated in various combinations, but not through a complete Cartesian product. Values in **bold** are from our submission.

languages	weights	unidir.	bidir.
de-dsb	full	70.0	70.5
de-{dsb,hsb}	full	75.7	75.8
de-{dsb,hsb}	LoRA	44.1	
de-hsb	full	(36.1)	(45.1)

Table 3: chrF scores for German to Lower Sorbian translations of *2025.dev.de-dsb* comparing unidirectional (into Sorbian {dsb,hsb}) vs. bidirectional (into Sorbian {dsb,hsb} and into German) corpora and different language pair combinations. These models are supervised finetuned (sft) from Qwen2.5-0.5B-Instruct. For completeness, in parentheses, we show Upper Sorbian systems' translations of Lower Sorbian.

languages	weights	unidir.	bidir.
de-dsb	full	(42.3)	(42.6)
de-{dsb,hsb}	full	80.2	80.2
de-{dsb,hsb}	LoRA	54.8	
de-hsb	full	79.2	79.2

Table 4: chrF scores for German to Upper Sorbian translations of *2025.dev.de-hsb* comparing unidirectional (into Sorbian {dsb,hsb}) vs. bidirectional (into Sorbian {dsb,hsb} and into German) corpora and different language pairs combinations. These models are supervised finetuned (sft) from Qwen2.5-0.5B-Instruct. For completeness, in parentheses, we show Lower Sorbian systems' translations of Upper Sorbian.

	Stopping Criterion		
model size	`eval_loss`	`#epochs`	`chrF`
0.5B	75.1	75.7	69.8
1.5B	76.7	77.7	74.7
3B	79.1	79.8	75.5

Table 5: chrF↑ scores for German to Lower Sorbian translations of *2025.dev.de-dsb* using different stopping criteria.

	Stopping Criterion		
model size	`eval_loss`	`#epochs`	`chrF`
0.5B	80.0	80.2	77.0
1.5B	81.8	82.2	80.9
3B	83.0	83.5	81.6

Table 6: chrF↑ scores for German to Upper Sorbian translations of *2025.dev.de-hsb* using different stopping criteria.

chrF↑	template		
Languages	chatml	empty	qwen
de-dsb	70.1	70.0	70.0
de-{dsb,hsb}	75.7	75.7	75.6
de-hsb	(40.6)	(36.1)	(40.7)

Table 7: chrF↑ scores for German to Lower Sorbian on *2025.dev.de-dsb* using different templates. These models are supervised finetuned(sft) from Qwen2.5-0.5B-Instruct. For completeness, in parentheses, we show Upper Sorbian systems' translations of Lower Sorbian.

chrF↑	template		
Languages	chatml	empty	qwen
de-dsb	(42.4)	(42.3)	(42.4)
de-{dsb,hsb}	80.2	80.2	80.1
de-hsb	79.4	79.2	79.3

Table 8: chrF↑ scores for German to Upper Sorbian on *2025.dev.de-hsb* using different templates. These models are supervised finetuned(sft) from Qwen2.5-0.5B-Instruct. For completeness, in parentheses, we show Lower Sorbian systems' translations of Upper Sorbian.

based), `chatml` (a generic chat template), and `empty` (a non-chat template). Figure 1 shows example input for the `empty` template. Figure 2 shows example input for the `chatml` and `qwen` template. The `qwen` and `chatml` templates, given the same input, render the same output except that `qwen`'s output has an additional system prompt. The system prompt is: "*You are Qwen, created by Alibaba Cloud. You are a helpful assistant*". Figure 4 shows a rendered example, using `qwen`'s template, as seen at training time by `LLaMa-Factory`.

We trained 9 models, the Cartesian product of three templates with three language corpora, using the identical configurations except for the datasets and the template. The models were trained to translate from German into Upper Sorbian and/or Lower Sorbian. We then translated the *2025.dev.de-dsb* and *2025.dev.de-hsb* sets using `LLaMa-Factory`. Table 7 shows chrF scores for translations into Lower Sorbian and Table 8 shows chrF scores for translation into Upper Sorbian when training a system on a given set of languages and template.[10] Naturally, we would not expect that a model trained on de-dsb be particularly good at translating Upper Sorbian or a de-hsb model at translating Lower Sorbian. This is confirmed in Table 7 and Table 8. Otherwise, the template choice does not seem to substantially impact the translation quality. The earlier translation direction and corpora experiments as well as the stopping criterion experiments were performed using the `empty` template.

4 Question Answering

We used our MT-finetuned models as the base models from which to train for the QA task. Our main experimental method for Multiple Choice Question Answering (MCQA) training was to use Direct Preference Optimization (Rafailov et al., 2024). We also ran limited experiments using supervised finetuning, but initial results showed a large drop in performance on the translation task as a result.

4.1 Data

We used all the available QA data for Upper Sorbian and Lower Sorbian. In Section 4.1.1 we describe how we split the data into *dev* and *train*. Section 4.1.2 discusses mitigation for potential position bias and Section 4.1.3 describes an approach to augmenting the data to try to better handle the range of possible response list sizes.

4.1.1 Dev/Train Split

To track QA performance during training, we split the QA corpora provided by the organizers[11] into 20% *dev* set and 80% *train*. Sampling of *dev* was performed using an unweighted reservoir sampling[12] algorithm. Table 11 shows which document ids were used in our *dev* set.

4.1.2 Position Bias

Position bias, where the position of answers influences the model's accuracy in answering, is known to be a problem for LLMs (Pezeshkpour and Hruschka, 2024, i.a.) and this problem can be exacerbated if there is also bias in the training data (i.e., if the correct answer is frequently shown in the same position among the options). To prevent answer position bias, we did data augmentation by adding repeated versions of the questions after permuting the answer order. We generate up to a maximum of 16 permutations per question-answer pair. That is,

[10] For completeness, we show Upper Sorbian systems' translations of Lower Sorbian text as well as the reverse, but show them in parentheses.

[11] https://github.com/TUM-NLP/llms-limited-resources2025

[12] https://en.wikipedia.org/wiki/Reservoir_sampling

```
{
  "instruction": "Translate the following German text to Lower
   ↪  Sorbian. Put it in this format <dsb> Lower Sorbian translation
   ↪  </dsb>.",
  "input": "<deu> Sogar Franz, der überhaupt nicht gerne singt,
   ↪  brummt laut mit. </deu>",
  "output": "<dsb> Samo Franc, kenž zewšym rad njespiwa, barcy
   ↪  głosnje sobu. </dsb>"
}
```

Figure 1: Input example for `empty` template

```
{
  "conversations": [
    {
      "from": "human",
      "value": "Translate the following German text to Lower Sorbian.
       ↪  Put it in this format <dsb> Lower Sorbian translation
       ↪  </dsb>.\n<deu> Sogar Franz, der überhaupt nicht gerne
       ↪  singt, brummt laut mit. </deu>"
    },
    {
      "from": "gpt",
      "value": "<dsb> Samo Franc, kenž zewšym rad njespiwa, barcy
       ↪  głosnje sobu. </dsb>"
    }
  ]
}
```

Figure 2: Input example for `chatml`/`qwen` template

we generate the minimum between the factorial of the number of possible answer orders for a given question or 16.

4.1.3 Augmentation

Noting that `lm-eval` considers all 16 possible answer positions and that many of the questions did not have a total of 16 answers, we augmented the answer sets during training such that each augmented QA instance had the correct answer paired with an incorrect answer in one of the other 15 answer positions. This was done after the position bias augmentation. Direct preference optimization then contrasts the correct and rejected answer during training. Table 9 shows the final number of questions in each QA corpora. Doing this augmentation after the position bias augmentation may have undermined the effectiveness of that position bias mitigation approach, as this second augmentation may reinforce a bias towards the first few answers being the correct response.

	original		augmented	
Question	*dev*	*train*	*dev*	*train*
DSB_A1	6	24	180	720
DSB_A2	5	23	210	1350
DSB_B1	8	36	1215	5250
DSB_B2	11	45	1695	7830
HSB_A1	6	24	180	720
HSB_A2	5	23	210	1350
HSB_B1	8	36	1215	5250
HSB_B2	11	45	1695	7830

Table 9: Number of questions per question type.

4.2 Training

To train our models for the QA task, we once again used `LLaMa-Factory` to tune all weights but used direct preference optimization (called the

	chrF↑		accuracy	
	dsb	hsb	dsb-qa	hsb-qa
TartuNLP	78.20	86.33	**57.56**	**58.10**
NRC	**78.24**	**87.20**	32.20	29.05
SDKM	64.34	75.73	51.71	55.24
baseline	12.21	13.88	45.85	42.86

Table 10: Official results.

`dpo stage` in `LLaMa-Factory`) rather than supervised finetuning. We explored learning rates of `7.0e-05`, `2.0e-05`, `7.0e-06`, `2.0e-06`, `2.0e-07` and `1`, `3` or `5` training epochs. The gradient norm was set to `1.0`, the warmup ratio to `0.01`, weight decay to `0.01`. The learning rate scheduler was set to `inverse_sqrt` and the optimizer was `adamw_torch`. The effective batch size was `32` because we used a per device batch size of `1` and `8` gradient accumulation steps over `4` GPUs. We encountered challenges in training a 3B parameter model for QA, primarily due to insufficient GPU memory to accommodate even a batch size of one. Consequently, we were restricted to smaller model sizes, specifically the 0.5B and 1.5B parameter variants. The inability to train a 3B parameter model for QA, coupled with the competition's requirement to employ a singular model for both MT and QA, rendered any further attempts to train a 3B parameter model for MT futile. Figure 6 and Figure 7 illustrate a correct answer and a rejected answer respectively, rendered from the corpus sample in Figure 5 the QA template.

4.3 Results

We evaluated on our QA *dev* sets and saw improvements over random chance. However, later examination of lower-than-expected scores when evaluating on the training data indicated that the models are not training as well we would have hoped. Table 14 shows our *dev* and *train* accuracies of our submission system compared to random chance, highlighting that our system was not effectively learning.

We are continuing to explore the reasons behind this, whether it relates to the data augmentation approach, the data splits for training, the fact that our *dev* was quite small, the selection of hyperparameters, the choice to use DPO, or some combination thereof.

5 Submissions

Here, we describe the specifics of our primary submissions. We submitted a single model that was trained to translate into both Upper Sorbian and Lower Sorbian. Note that our translations were obtained from `LLaMa-Factory` for inference while the QA output are produced using `lm-eval`.[13]

At submission time our best model attempting to balance between MT and QA performance, according to our *dev* sets for MT and QA, was based on `Qwen2.5-1.5B-Instruct`[14]. It used `LLaMa-Factory`'s supervised finetuning (`sft training stage`) and trained all weights (not LoRA). Using the `chatml` template, it trained on our filtered corpora from the top part of Table 1: 855,409 training examples as reported in the log. The learning rate was `7e-5` with a learning rate scheduler of type `cosine`. We capped the maximum gradient norm of `1.0`, and used no warmup and no weight decay. We used `8` gradient accumulation steps, a per device training batch size of `8` on `4` GPUs for a total training batch size of `256`. As this was a model that we trained near the start of the task, not knowing what would be a good number of training epochs given the number of training examples, we arbitrarily used `20` epochs (66,820 updates) and left the model to train until completion, expecting that the training loss curve would guide us in selecting a smaller number of epochs for future training sessions. Over the entire training, this model has seen 2,941,371,584 input tokens and required 2.31e+19 Flops of computation.

The official results for the task are shown in Table 10. As might be expected for MT-only systems, we have the highest performance on the MT tasks, but the lowest performance on QA, even falling below random chance for HSB-QA. This was in contrast to the performance that we observed on development data (where we did improve on random chance, even for MT-only systems), which merits additional analysis. Our leading hypotheses are that our *dev* set was too small and that DPO was not suited for the QA task.

We note that, after the late decision to use an MT-only system, we might have benefited from

[13] Our removal of newlines for the MT training data meant that we did not typically meet the expected stopping criteria for `lm-eval`, hence our use of `LLaMa-Factory` for inference.

[14] https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct

selecting a 3B parameter model rather than a 1.5B parameter model, as we did observe increasing MT performance with increasing model size in early stopping criterion experiments. While our stopping criterion experiments showed that chrF scores increased as the number of model parameters increased, with 3B outperforming 1.5B, we had paused our work on 3B parameter models in order to focus on the QA experiments. Had we run additional MT training (in line with our final setup the submission) on 3B model parameters, we expect they would have outperformed our submitted 1.5B parameter model results. We opted not to submit an early 3B MT-only model.

6 Conclusion

We submitted systems for the Upper Sorbian and Lower Sorbian portions of the shared task. Due to the lack of success of our approaches to QA, we submitted systems that were trained only on MT. As expected, our systems performed well on translation (our main focus) but underperformed on QA. If we had examined the QA accuracies on our training data earlier, as shown in Table 14, we could have concluded that our training for QA was ineffective. We were misled by the performance on the development set, which may have been too small for the task; larger development sets or cross-validation approaches could have also alerted us to these issues earlier. For MT for these low-resource languages, we observed benefits of training using both Lower Sorbian and Upper Sorbian parallel text (on the order of 5 chrF points for Lower Sorbian and 1 chrF point for Upper Sorbian).

We had observed some small improvements in QA even when training only on MT when evaluated on *dev* data, but did not find the same results on the test set. We continue to explore the reasons for this.

Limitations

We focused primarily on the machine translation portion of the task. There remain areas to explore in more detail, such as the impacts of model size (number of parameters) on performance. For some of our experiments, such as comparing unidirection and bidirectional training data, we trained the smallest (0.5B parameter) models; it is not known how well these results will generalize as the number of model parameters scales. There remains additional work to do on the QA task and on analyzing why our training for that did not perform as expected.

Acknowledgements

We would like to thank the reviewers for their suggestions and comments and the shared task organizers for their work on the task.

References

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Data Cleaning

Figure 3 shows the regular expression used to perform clean up of the corpora.

B Dev Question IDs

Table 11 lists the QA question ids for items in the *dev* set.

Question	ids
DSB_A1	A1.1.H02, A1.1.H11, A1.1.H7, A1.1.L14, A1.1.L6, A1.1.L7
DSB_A2	A2.1.H02, A2.1.H11, A2.1.L12, A2.1.L14, A2.1.L3
DSB_B1	B1.1.S12, B1.1.S13, B1.1.S19, B1.1.S23, B1.1.S3, B1.1.S5, B1.1.S7, B1.1.S9
DSB_B2	B2.1.H17, B2.1.H3, B2.1.H5, B2.1.L14, B2.1.L8, B2.1.S18, B2.1.S19, B2.1.S22, B2.1.S23, B2.1.S7, B2.1.S8
HSB_A1	A1.1.H02, A1.1.H11, A1.1.H7, A1.1.L14, A1.1.L6, A1.1.L7
HSB_A2	A2.1.H02, A2.1.H11, A2.1.L12, A2.1.L14, A2.1.L3
HSB_B1	B1.1.S12, B1.1.S13, B1.1.S19, B1.1.S23, B1.1.S3, B1.1.S5, B1.1.S7, B1.1.S9
HSB_B2	B2.1.H17, B2.1.H3, B2.1.H5, B2.1.L14, B2.1.L8, B2.1.S18, B2.1.S19, B2.1.S22, B2.1.S23, B2.1.S7, B2.1.S8

Table 11: Questions ids for *dev* sets.

```
s/[\x01-\x09\x0B\x0C\x0E-\x1D\x7F]//g;
s/\x0D//g;
s/\t/ /g;
s|\\\\ ?[rn]| |g;
s|\\ ?[rn]| |g;
s/  */ /g;
s/  *$//;
```

Figure 3: Regular expressions used to clean up input corpora.

C Random Baselines

Table 12 lists the random chance baseline for selecting correct responses on the *dev* set, while Table 13 shows the same for the test set.

Question	DSB(%)	HSB(%)
A1	50.00	50.00
A2	42.86	42.86
B1	25.19	25.19
B2	21.74	21.74
Overall	31.81	31.81

Table 12: *dev* set random chance baseline accuracy.

Question	DSB(%)	HSB(%)
A1	50.00	50.00
A2	42.86	42.86
B1	25.19	25.19
B2	21.74	21.74
C1	32.71	31.15
Overall	31.93	31.65

Table 13: Test set random chance baseline accuracy.

D Supervised Finetuning

Figure 4 shows Figure 2 rendered using Qwen's template.

E Question Answering

Figure 5 illustrates a training example for QA. Figure 6 is the resulting output of a correct answer whereas Figure 7 is the resulting output of a rejected answer as seen at training time by the model.

F Submission QA Accuracies

Table 14 shows random chance accuracies alongside the accuracies of our submitted system on both *dev* and *train*, emphasizing the inadequate training.

Source

```
<|im_start|>system
You are Qwen, created by Alibaba Cloud. You are a helpful
 ↪  assistant.<|im_end|>
<|im_start|>user
Translate the following German text to Lower Sorbian. Put it in this
 ↪  format <dsb> Lower Sorbian translation </dsb>.
<deu> Sogar Franz, der überhaupt nicht gerne singt, brummt laut mit.
 ↪  </deu><|im_end|>
<|im_start|>assistant
<dsb> Samo Franc, kenž zewšym rad njespiwa, barcy głosnje sobu.
 ↪  </dsb><|im_end|>
```

Target

```
<dsb> Samo Franc, kenž zewšym rad njespiwa, barcy głosnje sobu.
 ↪  </dsb><|im_end|>
```

Figure 4: Rendered `qwen` template for MT.

```
{
  "instruction": "žona: Pětš, psiźoš sobu do kina? \nPětš: Halo,
   ↪  Monika. Njewěm hyšći. Ga? \nžona: Pónjeźele? \nPětš: Derje,
   ↪  pónjeźele.\n\nQuestion:\nGa cotej Pětš a Monika do kina
   ↪  hyś?\n\nPossible answers:\n1 pónjeźele\n2 pětk\n\nAnswer:",
  "chosen": "1",
  "rejected": "2"
}
```

Figure 5: Input example for Question Answering using DPO template

```
žona: Pětš, psiźoš sobu do kina?
Pětš: Halo, Monika. Njewěm hyšći. Ga?
žona: Pónjeźele?
Pětš: Derje, pónjeźele.

Question:
Ga cotej Pětš a Monika do kina hyś?

Possible answers:
1 pónjeźele
2 pětk

Answer:1
```

Figure 6: Rendered QA template for the chosen answer

```
žona: Pětš, psiźoš sobu do kina?
Pětš: Halo, Monika. Njewěm hyšći. Ga?
žona: Pónjeźele?
Pětš: Derje, pónjeźele.

Question:
Ga cotej Pětš a Monika do kina hyś?

Possible answers:
1 pónjeźele
2 pětk

Answer:2
```

Figure 7: Rendered QA template for the rejected answer

	DSB					HSB				
Submission	Sorbian	A1	A2	B1	B2	Sorbian	A1	A2	B1	B2
dev	45.13	66.67	40.00	37.50	36.36	40.59	66.67	40.00	37.50	18.18
train	26.87	45,83	26.09	22.22	13.33	32.84	50.00	39.13	22.22	20.00
Random Chance	31.81	50.00	42.86	25.19	21.74	31.81	50.00	42.86	25.19	21.74

Table 14: Question Answering *dev*/*train* accuracies (%) for our submission.

TartuNLP at WMT25 LLMs with Limited Resources for Slavic Languages Shared Task

Taido Purason and Mark Fishel
Institute of Computer Science
University of Tartu, Estonia
{taido.purason, mark.fisel}@ut.ee

Abstract

This paper describes the TartuNLP submission to the Upper Sorbian (*hsb*) and Lower Sorbian (*dsb*) tracks of the WMT25 LLMs with Limited Resources for Slavic Languages shared task, which jointly targets machine translation (MT) and question answering (QA). We develop a single multilingual model based on `Qwen2.5-3B-Instruct` by continuing pretraining on Sorbian monolingual and parallel data together with general instruction datasets, combining language acquisition and instruction-following in a single step. The resulting model delivers substantial improvements over the baseline `Qwen2.5-3B-Instruct` model and also achieves the highest ranking for both tasks in the *hsb* and *dsb* shared task tracks.

	DE-HSB		HSB-QA		final
Team	**chrF++**	**points**	**acc**	**points**	**points**
TartuNLP	86.33	4	58.10	4	8
NRC	87.20	4	29.05	1	5
SDKM	75.73	2	55.24	3	5
baseline	13.88	1	42.86	2	3

Table 1: **Upper Sorbian** (hsb) rankings.

	DE-DSB		DSB-QA		final
Team	**chrF++**	**points**	**acc**	**points**	**points**
TartuNLP	78.20	4	57.56	4	8
NRC	78.24	4	32.20	1	5
SDKM	64.34	2	51.71	3	5
baseline	12.21	1	45.85	2	3

Table 2: **Lower Sorbian** (dsb) rankings.

1 Introduction

This paper presents an overview of the TartuNLP systems developed for the WMT25 Limited Resource Slavic Languages shared task (Okabe et al., 2025). This shared task aimed to create a single large language model (LLM) capable of jointly performing both machine translation (MT) and question answering (QA) for less-resourced Slavic languages. The participants of the shared tasks were limited to using the Qwen2.5 model family (Qwen Team, 2024) with a size constraint of 3B parameters. Our team participated in the Upper Sorbian (*hsb*) and Lower Sorbian (*dsb*) tracks, both endangered languages spoken by only about 20,000–30,000 people in total (Moseley, 2007). The taxonomy of Joshi et al. (2020) categorizes both languages as category-1, *the scraping-bys*. We focused on building a single model that supports both tasks and languages simultaneously. This joint objective introduces a specific challenge: while a moderate amount of MT data exists for Sorbian, there is no QA training data, requiring balancing performance across both tasks.

Although one or both of the Sorbian languages have been included in recent massively multilingual models (Imani et al., 2023; Lin et al., 2024; Ji et al., 2025b,a), to our knowledge, no prior work has developed dedicated LLMs for Sorbian.

We build on recent work adapting LLMs to extremely low-resource languages through continued pretraining and instruction tuning (Purason et al., 2025; Etxaniz et al., 2024; Sainz et al., 2025). Our approach follows Sainz et al. (2025), who demonstrated for the Basque language that combining language acquisition and instruction tuning in a single step and starting from an instruction-tuned model is beneficial. We also adopt this joint learning of instruction-following and language for an instruction-tuned base model in our system.

We continually pretrain `Qwen2.5-3B-Instruct` (Qwen Team, 2024) on a mix of monolingual documents and sentences in *hsb* and *dsb*, general instruction-following data (primarily in English), and Sorbian MT instructions. We supplement the data provided by the organizers with document-level texts from Fineweb-2 (Penedo et al., 2025) and Wikipedia articles (Wikimedia

Proceedings of the Tenth Conference on Machine Translation, pages 1143–1150

Foundation). Our final model outperforms the baseline `Qwen2.5-3B-Instruct` and also achieves the highest rank in the shared task (see Tables 1 and 2). **We publish the final model on HuggingFace**[1].

2 Datasets

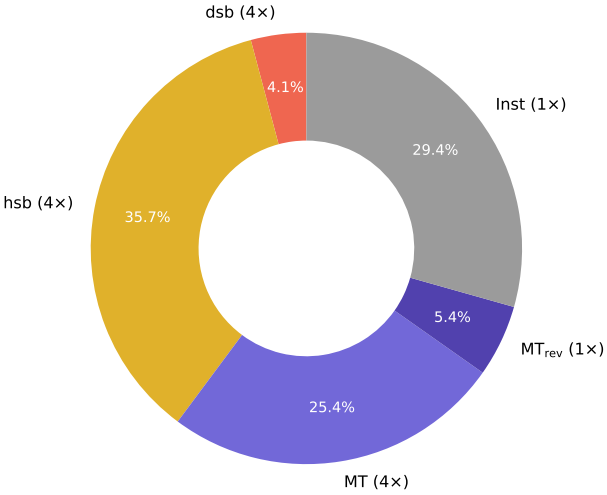


Figure 1: **Data mix** for the final model (% of total training tokens). The number of epochs for each dataset is stated in parentheses. The resulting training set is 1.198B tokens.

Dataset	Texts	Words	Chars
HSB			
Sentence-level (WMT)	1.8M	25.2M	166.2M
Document-level	53.4K	13.1M	90.9M
- Fineweb-2	40.2K	12.1M	83.6M
- Wikipedia	13.2K	1.1M	7.3M
Total (dedup)	1.7M	35.9M	240.6M
DSB			
Sentence-level (WMT)	170.5K	2.5M	15.3M
Document-level	9.5K	2.0M	13.9M
- Fineweb-2	6.3K	1.8M	12.2M
- Wikipedia	3.3K	249.5K	1.7M
Total (dedup)	169.8K	4.4M	28.4M

Table 3: Monolingual dataset statistics.

	de-hsb	de-dsb	dsb-hsb
Sentence pairs	636.3K	212.2K	62.6K

Table 4: The parallel data sentence pair counts.

Monolingual data (see Table 3). We used all of the monolingual data provided by the organizers for this year (Okabe et al., 2025) and the data from previous WMT Sorbian shared tasks (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022), which was sentence-level aligned. Additionally, we used Upper and Lower Sorbian Fineweb-2 (Penedo et al., 2025) documents and Upper and Lower Sorbian Wikipedia documents from the 2025 05 20 dump (Wikimedia Foundation) extracted with WikiExtractor (Attardi, 2015). We also experimented with using German Fineweb-2 (Penedo et al., 2025) documents for pretraining, however, we did not use this data for training the final submission model.

Parallel data (see Table 4). Our parallel data is entirely from this and previous years' shared tasks (Fraser, 2020; Libovický and Fraser, 2021; Weller-Di Marco and Fraser, 2022). The sentence pairs were formatted as instructions in a chat template. Since the task was to translate from German to the Sorbian languages, we trained on translation into the Sorbian languages (de-hsb, de-dsb, hsb-dsb, dsb-hsb) for 4 epochs. We refer to this dataset as MT. In addition to that, we also added one epoch of data in the hsb-de and dsb-de directions as well (referred to as MT$_{rev}$).

The monolingual and parallel data was deduplicated with normalization from Stopes (Pierre Andrews, 2022; NLLB Team et al., 2022). We also removed the held-out and validation data using the same normalization method.

Instructions (see Table 11). We used instruction data from Magpie (Xu et al., 2025), Aya (Singh et al., 2024), EuroBlocks (Martins et al., 2025) OpenAssistant2 (Köpf et al., 2023), and FLAN v2 (Longpre et al., 2023). We removed instructions that were not in English, German, or the closest Slavic languages to the Sorbian languages. Still, 98.6% of instructions in the resulting dataset are in English.

The final mix. The data mix of the submitted model consisted of 1.198B tokens and is displayed in Figure 1.

3 Methodology

We used `Qwen2.5-3B-Instruct` (Qwen Team, 2024) as our base model, motivated by the findings of Sainz et al. (2025), who demonstrated that continued pretraining on already instruction-tuned models is effective for low-resource languages. Following their findings, we combined the language acquisition and instruction-tuning in the same continued pretraining step. Our continued pretraining

[1] huggingface.co/tartuNLP/Qwen2.5-3B-Instruct-hsb-dsb

Model							MT (BLEU / chrF++) de-dsb	MT (BLEU / chrF++) de-hsb	QA (acc) dsb	QA (acc) hsb	Evaluation setting
BASELINES											
`emma-500-llama3.1-8b-bi`							10.8 ± 0.7 / 35.3 ± 0.8	20.0 ± 1.1 / 47.2 ± 0.7	48.0 ± 7.4	41.2 ± 7.2	MT: 5-shot; QA: 3-shot
`Qwen2.5-3B-Instruct`							0.8 ± 0.2 / 12.6 ± 0.3	1.3 ± 0.2 / 17.4 ± 0.3	55.3 ± 7.8	58.0 ± 7.4	MT: 5-shot; QA: 3-shot
CONTINUED PRETRAINING											
	dsb	hsb	MT💬	MT_{rev}💬	Inst💬	deu					
(1)	4x	4x					19.9 ± 0.6 / 45.3 ± 0.5	28.4 ± 1.2 / 54.2 ± 0.6	69.0 ± 6.9	67.9 ± 6.9	MT: 5-shot; QA: 3-shot
(2)	4x	4x	4x				58.6 ± 1.0 / 77.1 ± 0.7	66.6 ± 0.7 / 82.1 ± 0.4	66.1 ± 7.1	70.9 ± 6.8	MT: 0-shot💬; QA: 3-shot
(3)	4x	4x	4x		1x		60.2 ± 1.2 / 78.1 ± 0.7	67.2 ± 0.7 / 82.6 ± 0.4	67.5 ± 7.2	**73.1 ± 6.6**	0-shot💬
(4)	4x	4x	4x	1x	1x		62.0 ± 1.0 / 79.2 ± 0.6	68.3 ± 0.7 / 83.2 ± 0.4	67.5 ± 7.2	70.4 ± 6.7	0-shot💬
(5)	4x	4x	4x	1x	1x	25%	62.1 ± 1.1 / 79.2 ± 0.6	68.6 ± 0.7 / 83.4 ± 0.4	65.7 ± 7.1	65.1 ± 6.9	0-shot💬
(4)	**Final submission**					validation	**62.8 ± 1.1 / 79.8 ± 0.6**	**69.1 ± 0.7 / 83.7 ± 0.4**	**69.3 ± 7.0**	69.3 ± 6.9	MT: 0-shot💬†; QA: 1-shot💬
						test	- / 78.20	- / 86.33	57.56	58.10	

Table 5: Validation set scores for the baselines and fine-tuned models. The shared task's final submission validation and test set scores. 💬 - chat instruction format; † - beam search decoding with beam size 4.

was performed jointly on monolingual documents and sentences, parallel MT data, and instruction-formatted data.

Our initial experiments indicated that the model begins to overfit on Sorbian data around the fourth epoch. A similar finding was also made in Purason et al. (2025), who observed overfitting at 4 epochs for low-resource languages of similar size. With this in mind, we repeated the Sorbian monolingual and parallel data four times during training. The instruction data was repeated once, while the significantly more abundant German monolingual data (not used for the final submission) was limited to 25% of the total training token budget. It should be noted that we did not conduct a thorough investigation into the different data sampling strategies and combinations. A different number of epochs or curriculum learning might provide better results.

For instruction-formatted samples, we applied loss only on the assistant (target) tokens. All datasets were packed into sequences of 4096 tokens, with any overflow tokens carried into the next training sequence. The hyperparameters for the model training are listed in Table 10. The models were trained on either 8 or 16 nodes, each consisting of 4 AMD MI250x GPUs (acting as 8 units) on the LUMI supercomputer. The training of the final model took 139 GPU-hours.

4 Evaluation

We evaluate our models using the `lm-eval-harness` framework (Gao et al., 2024), which we also use to generate the final shared task submissions. Evaluation is conducted using a few-shot and zero-shot prompting strategies, depending on the training strategy. We use a few-shot evaluation without a chat template and a zero-shot evaluation using a chat-style format.

For QA, we follow a multiple-choice setup, selecting the answer option with the highest log-probability from a predefined set of candidates. We calculate the accuracies and report the average across the language levels in the validation set.

For the MT evaluation, we calculated the BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) scores. We also report 95% confidence intervals calculated from standard errors reported by `lm-eval-harness`.

For the final submission, we apply beam search with a beam size of 4 for MT, and use one-shot prompting with the chat template for QA, where each example is treated as a separate turn in a multi-turn conversation. We use greedy decoding in all other evaluation settings (beam size of 1).

5 Results

5.1 Main results

Table 5 summarizes the performance of our models on both machine translation (MT) and question answering (QA) for Upper and Lower Sorbian. We compare our systems against two open-weight baselines: `emma-500-llama3.1-8b-bi` (Ji et al., 2025a), which includes Sorbian in its training data, and `Qwen2.5-3B-Instruct` (Qwen Team, 2024).

As expected, we observe that the `Qwen2.5-3B-Instruct` model performs poorly on Sorbian MT benchmarks before any continued pretraining. In contrast, `emma-500`, which was trained with Sorbian data, performs noticeably better in MT. However, the trend reverses for QA:

`Qwen2.5-3B-Instruct` significantly outperforms `emma-500`, highlighting the strength of instruction tuning for QA even in low-resource settings.

Our continued pretraining configurations substantially outperform the baselines across both MT and QA tasks, demonstrating the effectiveness of the language adaptation. From these results, we find that:

- Including MT data during continued pretraining yields large gains in translation quality compared to relying solely on few-shot prompting.
- Adding instruction-following data provides additional improvements for both tasks.
- Adding a small amount of reverse-direction (into German) MT data (MT_{rev}) appears to slightly boost MT performance without harming QA.
- Allocating 25% of the training budget to monolingual German data does not improve MT and slightly degrades QA.

Despite these trends, due to the small size of the QA evaluation set (n = 162) and the resulting wide confidence intervals, it remains difficult to draw definitive conclusions about which components contributed most to QA performance. Nevertheless, our results confirm the benefit of continued pretraining on task- and language-specific data, particularly when jointly targeting MT and QA with a single model.

Our final submission additionally used one-shot prompting for QA and beam search with a beam size of 4, which slightly increased the scores, although this increase is likely not significant.

Our submission achieved the highest rank in QA and MT tasks for both languages and was also the overall winner in those languages.

5.2 Combined or separate Sorbian training

Since they are closely related, we also investigate how much the Sorbian languages benefited from joint training. From the results in Table 6, we see that both languages benefit from the joint training, especially for the generative MT task. It is also apparent that for the QA task, the model trained on only one of the Sorbian languages can perform quite well for the other, even without the data, suggesting that we gain language understanding from the other Sorbian language.

	MT (BLEU)		**QA** (acc)	
Sorbian data	**de-dsb**	**de-hsb**	**dsb**	**hsb**
hsb + dsb	**60.2 ± 1.2**	**67.2 ± 0.7**	67.5 ± 7.2	**73.1 ± 6.6**
dsb	54.0 ± 1.0	9.7 ± 0.4	**69.6 ± 7.0**	67.3 ± 7.0
hsb	11.8 ± 0.6	65.6 ± 0.7	61.0 ± 7.4	71.8 ± 6.8

Table 6: MT (BLEU) and QA (acc) scores when training the Sorbian languages **together vs separately**. *hsb+dsb* configuration is equal to (3) in Table 5. Zero-shot prompting with a chat template was used.

5.3 Effect of the document-level data

		MT (BLEU)		**QA** (acc)	
Sorbian	**N_{chars}**	**de-dsb**	**de-hsb**	**dsb**	**hsb**
sent-level	181.5M	11.9 ± 1.2	19.8 ± 1.7	**71.3 ± 6.8**	71.3 ± 6.7
doc-level	104.8M	21.3 ± 0.9	29.3 ± 0.7	**71.3 ± 6.8**	70.4 ± 6.7
combined	269.1M	**24.4 ± 1.0**	**32.5 ± 0.7**	68.8 ± 6.8	**71.9 ± 6.6**

Table 7: Results for training with HSB and DSB data either **document level, sentence-level, or combined** (without MT examples). The training data includes 1 epoch of *INST* and 25% token budget for *deu* (doc-level). 5-shot BLEU scores are reported for MT, and 3-shot accuracy is reported for QA. N_{chars} is the number of characters in the *hsb* and *dsb* datasets

We examine the effect of incorporating document-level data in addition to sentence-level data. No explicit *hsb/dsb* MT training data was used in these experiments. The results in Table 7 indicate that document-level data is more beneficial than sentence-level data for MT when evaluating in a few-shot setting. This is somewhat surprising given that the sentence-level dataset is substantially larger, and the quality of the Fineweb-2 portion of the document-level dataset has not been verified. Nevertheless, combining document- and sentence-level datasets yields the highest MT scores. For QA, we did not observe significant differences between the data setups.

5.4 MT Supervised Fine-tuning

We instruction-tune our model as a separate step with machine-translation examples (4 epochs of *de-hsb*, *de-dsb*, *hsb-dsb*, and *dsb-hsb*) formatted in chat format as instructions, instead of training jointly. In Table 8, this approach shows a slight increase in the BLEU scores of the machine translation benchmarks. However, our model loses its capability to answer questions, so this strategy does not satisfy the goals of the shared task.

Continued pretraining						SFT_{MT} 💬	MT (BLEU)		QA (acc)	
dsb	hsb	MT 💬	MT_{rev} 💬	Inst 💬	deu		de-dsb	de-hsb	dsb	hsb
4x	4x	4x	1x	1x	25%	-	62.1 ± 1.1	68.6 ± 0.7	**65.7 ± 7.1**	**65.1 ± 6.9**
4x	4x			1x	25%	4x	**64.8 ± 1.0**	**70.6 ± 0.6**	36.6 ± 7.1	34.7 ± 7.2

Table 8: **Machine translation SFT** results (validation set) evaluated with a zero-shot chat format. 💬 - conversational instruction format.

	MT (BLEU)		QA (acc)	
Model	**de-dsb**	**de-hsb**	**dsb**	**hsb**
CPT	**62.1 ± 1.1**	**68.6 ± 0.7**	65.7 ± 7.1	65.1 ± 6.9
CPT + 0.1 BASE	61.8 ± 1.1	68.4 ± 0.7	66.6 ± 7.1	67.1 ± 7.0
CPT + 0.3 BASE	58.1 ± 1.1	65.8 ± 0.7	**67.9 ± 7.0**	**71.1 ± 6.8**

Table 9: **SLERP-merged models'** zero-shot chat-formatted validation set scores. $+x$ BASE means that the merging weights of the base model (`Qwen2.5-3B-Instruct`) are x while the Sorbian continually pretrained (CPT) model weight stays 1.0

5.5 Merging

Inspired by TowerLLM (Rei et al., 2025), who reported that merging with the original model improved general results while not harming translation significantly, we also decided to explore merging. SLERP merging with `mergekit` (Goddard et al., 2025) did show a slight increase in the QA scores (see Table 9), although its significance is questionable due to the small size of the validation set. We also noticed that merging started harming the translation quality at 0.3 weight for the base instruction-tuned model. With this in mind, we did not use it for the final model. It is possible that the benefit would be more apparent on other tasks that we did not measure due to the lack of validation data in the Sorbian languages.

6 Conclusion

We presented the TartuNLP submission to the WMT25 Shared Task on Limited Resource Slavic Languages, targeting both machine translation and question answering for Upper and Lower Sorbian. Our approach combined continued pretraining and instruction tuning in a single step, leveraging the Qwen2.5-3B-Instruct model. By integrating task-specific and monolingual Sorbian data, we achieved significant improvements over existing baselines and obtained the highest rank for both tasks in both languages. Our results demonstrate that a unified approach can effectively serve multiple low-resource language tasks, even under resource constraints.

7 Limitations

Our findings are limited by the fact that we only consider two tasks. Also, the MT data seems to have short sentence-level data, limiting the conclusions further. Since the test set of the QA task is relatively small, we have rather low confidence in the minor differences in scores of the approaches. We did not thoroughly explore all the design choices made for this system, and we did not explore data filtering, which could significantly impact the resulting model.

Acknowledgments

This work was supported by the Estonian Research Council grant PRG2006 (Language Technology for Low-Resource Finno-Ugric Languages and Dialects). All computations were performed on the LUMI Supercomputer through the University of Tartu's HPC center.

References

Giuseppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Julen Etxaniz, Oscar Sainz, Naiara Miguel, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for Basque. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14952–14972, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Fraser. 2020. Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2025. Arcee's mergekit: A toolkit for merging large language models. *Preprint*, arXiv:2403.13257.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Indraneil Paul, Hengyu Luo, and Jörg Tiedemann. 2025a. Massively multilingual adaptation of large language models using bilingual translation data. *Preprint*, arXiv:2506.00469.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2025b. Emma-500: Enhancing massively multilingual adaptation of large language models. *Preprint*, arXiv:2409.17892.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, and 1 others. 2023. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *Preprint*, arXiv:2401.13303.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and 1 others. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

Pedro Henrique Martins, João Alves, Patrick Fernandes, , Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Christopher Moseley. 2007. *Encyclopedia of the world's endangered languages*. Routledge.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Shu Okabe, Daryna Dementieva, Marion Di Marco, Lukas Edman, Kathy Hämmerl, Marko Měškank, Anita Hendrichowa, and Alexander Fraser. 2025. Findings of the WMT 2025 shared task for LLMs with limited resources for slavic languages: MT and QA. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language. *Preprint*, arXiv:2506.20920.

Kevin Heffernan Onur Çelebi Anna Sun Ammar Kamran Yingzhe Guo Alexandre Mourachko Holger Schwenk Angela Fan Pierre Andrews, Guillaume Wenzek. 2022. stopes - modular machine translation pipelines. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Taido Purason, Hele-Andra Kuulmets, and Mark Fishel. 2025. LLMs for extremely low-resource Finno-Ugric languages. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6677–6697, Albuquerque, New Mexico. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André

F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *Preprint*, arXiv:2506.17080.

Oscar Sainz, Naiara Perez, Julen Etxaniz, Joseba Fernandez de Landa, Itziar Aldabe, Iker García-Ferrero, Aimar Zabala, Ekhi Azurmendi, German Rigau, Eneko Agirre, and 1 others. 2025. Instructing large language models for low-resource languages: A systematic study for basque. *arXiv preprint arXiv:2506.07597*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Marion Weller-Di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2025. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *The Thirteenth International Conference on Learning Representations*.

A Hyperparameters

The training hyperparameters for the submitted model are in Table 10.

B Evaluation prompts

Evaluation prompts are presented in Figure 2.

C Instruction-tuning dataset overview

The full overview of the instruction data composition is presented in Table 11.

Hyperparameter	Value
Learning rate	1e-4
Optimizer	`AdamW`
Adam ϵ	1e-8
Adam β_1, β_2	0.9, 0.95
Sequence length	4096
Weight decay	0.1
Scheduler	`warmup-stable-decay`
Warmup steps	256
Decay steps	768
FSDP Strategy	`SHARD_GRAD_OP`
GPUs	64
Precision	`bfloat16`
Batch size (total)	128
Batch size (tokens)	524288
Training steps	2285
Training tokens	1,197,871,104

Table 10: Hyperparameters for the training of the submitted model.

MT

```
Translate the text from German to Lower Sorbian.\n\nGerman: {{de}}\nLower Sorbian:
```

MT (chat)

SYSTEM:

```
You are are a professional translator. Translate the following text from German to Lower Sorbian. Answer with the translated text.
```

USER:

```
{{de}}
```

QA

```
{context}\n\nQuestion:\n{question}\n\nPossible answers:\nA {answer_1}\nB {answer_2}...\nZ {answer_n}\n\nAnswer:
```

Figure 2: Prompts used for evaluation with `lm-eval-harness` (Gao et al., 2024).

Dataset		eng	pol	deu	ces	slk	slv	Total
CohereLabs/aya_dataset	Singh et al. (2024)	3944	1483	241	0	0	0	5668
Magpie-Align/Magpie-Llama-3.1-Pro-MT-300K-Filtered	Xu et al. (2025)	295830	36	228	21	3	3	296121
OpenAssistant/oasst2	Köpf et al. (2023)	22311	155	1785	5	0	0	24256
ai2-adapt-dev/flan_v2_converted	Longpre et al. (2023)	89982	0	0	0	0	0	89982
utter-project/EuroBlocks-SFT-Synthetic-1124†	Martins et al. (2025)	3057	916	1019	0	0	0	4992
Total		415124	2590	3273	26	3	3	421019

Table 11: Overview of the **instruction datasets**. †- *multilingual-synthetic-mmlu*, *synthetic-eurollm-9B*, and *multilingual-synthetic-arc* subsets from EuroBlocks.

JGU Mainz's Submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA

Hossain Shaikh Saadi,[1] **Minh Duc Bui,**[1]
Mario Sanz-Guerrero,[1] and **Katharina von der Wense**[1,2]
[1]Johannes Gutenberg University Mainz, Germany
[2]University of Colorado Boulder, USA

Abstract

This paper presents the JGU Mainz submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: Machine Translation and Question Answering, focusing on Ukrainian, Upper Sorbian, and Lower Sorbian. For each language, we jointly finetune a Qwen2.5-3B-Instruct model for both tasks with parameter-efficient finetuning. Our pipeline integrates additional translation and multiple-choice question answering (QA) data. For Ukrainian QA, we further use retrieval-augmented generation. We also apply ensembling for QA in Upper and Lower Sorbian. Experiments show that our models outperform the baseline on both tasks.

1 Introduction

While large language models (LLMs) are strong multitask learners for high-resource languages such as English, this is not the case for smaller LLMs and languages with limited data. In this setting, a trade-off presents between the performance on different tasks. The *WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages: MT and QA*[1] focuses on the development of relatively small LLMs ($\leq$3B parameters) that are capable of performing both machine translation (MT) and multiple-choice question answering (QA) in Slavic languages with limited amounts of data. Three languages – a mid-resource language, Ukrainian (UK), and two severely low-resource languages, Upper Sorbian (HSB) and Lower Sorbian (DSB) – are targeted, and only Qwen2.5 models with 0.5B, 1.5B, or 3B parameters are permitted. The MT source languages are Czech (CS) and English (EN) for Ukrainian as well as German (DE) for DSB and HSB.

Our proposed approach consists of a Qwen2.5-3B-Instruct model (Qwen et al., 2025), which is inherently multilingual and which we jointly finetune on both MT and QA data, combining the provided resources with additional datasets we curate ourselves. For DSB and HSB MT, we enhance the training data using synthetic data generated through back-translation. We further add an additional parallel dataset for DSB. For QA, we add a total of 16 high-quality English MCQ datasets. All QA datasets are enhanced via automatic translation such that they are bilingual (English and the target language). For Ukrainian QA, we incorporate retrieval-augmented generation (RAG) using domain-relevant Wikipedia pages and 10 books related to the subjects of the provided Ukrainian MCQ dataset.

At inference time, we use similarity-based few-shot in-context learning (ICL) for MT. For QA, we permute the order of the answer options, and average the probabilities for all options, to increase robustness against answer ordering biases (Pezeshkpour and Hruschka, 2024).

While not winning for any task–language combination, our primary submission consistently outperforms the baseline on all tasks, demonstrating the effectiveness of our approach. For DE–DSB translation, ChrF++ improves by over 55 points, while DE–HSB translation sees gains of over 65 points, reflecting substantial quality improvements. On DSB QA, accuracy increases by up to 12.34 percentage points, and, on HSB QA, accuracy increases by 10.27 points. For CS–UK and EN–UK MT, ChrF++ improves by 4.61 and 2.7, respectively, while, on Ukrainian QA, our submission outperforms the baseline by 4.66 accuracy points.

2 Data

2.1 Provided Data

Ukrainian For EN–UK and CS–UK MT, no training data are provided. Only development sets are given, containing 6,263 CS–UK and 5,108 EN–

[1] https://www2.statmt.org/wmt25/limited-resources-slavic-llm.html

Proceedings of the Tenth Conference on Machine Translation, pages 1151–1157
November 8-9, 2025

UK parallel sentences.

The UK QA data are curated from the External Independent Evaluation (3HO/ZNO), an exam for admission to Ukrainian universities. The dataset comprises a training set of 2,450 questions, a development set of 613 questions, and a test set of 751 questions. These questions cover three topics – Ukrainian History, Ukrainian Language, and Ukrainian Literature – and test both domain knowledge as well as reading comprehension.

Upper and Lower Sorbian MT For DE-to-DSB MT, 171k translation pairs are provided as training data. In addition, a 4,000-pair development set is also provided for system validation and evaluation. Some monolingual sentences, approximately 10k, are also provided along with the translation data.

For DE-to-HSB MT, 187k training translation pairs and a 4,000-pair development set are provided. 300k monolingual sentences are further available for model enhancement.

For both DSB and HSB, MCQ datasets are curated by the Witaj-Sprachzentrum. The questions are similar to the CEFR framework (A1 to C1), which follows the language certificate examinations. The difficulty of the questions ranges from simple true/false to complex multiple-choice formats with two to sixteen answer options. The development set contains 158 A1 to B2 level questions, while the test set for both languages has 205 questions for A1 to C1. An overview of the provided datasets is shown in Table 1.

2.2 Additional Data

Upper and Lower Sorbian MT In order to enhance our DE–DSB/HSB translation systems, we incorporate additional parallel data. For DSB, we translate the provided 10k monolingual examples into German. Then we create 10k additional translation pairs using the translated German sentences and the monolingual DSB sentences. Since 10k is a small amount, we additionally use 24k DE–DSB sentences from the Tatoeba bilingual dataset.[2] In turn, for HSB, we translate the first 100k monolingual sentences from the provided 300k sentences into German and create 100k additional translation pairs. Due to the substantial time required for translation, we translate only 100k sentences. To translate the DSB and HSB sentences into German, we first finetune two separate Qwen2.5-3B-Instruct models, one for HSB–DE and one for DSB–DE, using the respective provided parallel translation datasets. These models are finetuned by applying LoRA on all projection layers of the model.

Upper Sorbian, Lower Sorbian, and Ukrainian QA For QA in DSB, HSB, and Ukrainian, we select 16 English MCQA datasets, namely: (English) Global MMLU (Singh et al., 2025), CommonsenseQA (Talmor et al., 2018), ARC (Clark et al., 2018), Race (Lai et al., 2017), Dream (Sun et al., 2019), PIQA (Bisk et al., 2019), HellaSwag (Zellers et al., 2019), SCIQ (Johannes Welbl, 2017), MedMCQA (Pal et al., 2022), LogicQA (Liu et al., 2020), Quail (Rogers et al., 2020), SocialIQA (Sap et al., 2019), CosmosQA (Huang et al., 2019), OpenbookQA (Mihaylov et al., 2018), QASC (Khot et al., 2020), BoolQ (Clark et al., 2019). From these datasets, we sample up to 10k questions from each available split (training, development, and test), resulting in approximately 200k English MCQs. In order to translate the English MCQs into DSB and HSB, we first use `googletrans`[3] to translate the German sentences of the provided DE–DSB and DE–HSB translation examples into English, in order to create English–DSB/HSB translation pairs. Second, using this data, we finetune two separate Qwen2.5-3B-Instruct models on EN–DSB and EN–HSB MT. We use these two models to translate the English MCQs into DSB and HSB. For Ukrainian, we also translate the 200k English MCQs using `googletrans` directly, since Ukrainian is supported by Google Translate.

Ukrainian MT For CS–UK MT, we collect training data from OpenSubtitles (OPUS, 2003; Lison and Tiedemann, 2016), NeuTED (Qi et al., 2018), KDE4 (OPUS, 2003), and ELRC UKR Acts (ELRC, 2022). For EN–UK, we use OpenSubtitles (OPUS, 2003; Lison and Tiedemann, 2016), NeuTED (Qi et al., 2018), ELRC UKR Acts (Qi et al., 2018), and Multi30k (Saichyshyna et al., 2023). Across these sources, there are nearly 7 million sentence pairs per direction.

To reduce the number of sentence pairs, we apply a similarity-based retrieval method, using the provided development sets for CS–UK and EN–UK as the reference datasets. We embed each Ukrainian sentence from this dataset by performing mean-pooling over the last hidden states of Qwen2.5-3B-Instruct token outputs. For each sentence of

[2] https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/data/README-v2023-09-26.md

[3] https://github.com/ssut/py-googletrans

Task	Training Data	Dev Set	Test Set	Notes
EN→UK Translation	None	5,108 pairs	—	Only dev set available
CS→UK Translation	None	6,263 pairs	—	Only dev set available
Ukrainian QA (MCQs)	2,450 questions	613 questions	751 questions	From ZNO exam; covers History, Language & Literature; tests knowledge and comprehension
DE→DSB Translation	171k pairs	4,000 pairs	—	∼10k monolingual sentences provided
DE→HSB Translation	187k pairs	4,000 pairs	—	∼300k monolingual sentences provided
DSB QA (MCQs)	—	158 A1–B2 questions	205 A1–C1 questions	From Witaj-Sprachzentrum; CEFR-style; difficulty from true/false to multiple-choice (2–16 options)
HSB QA (MCQs)	—	158 A1–B2 questions	205 A1–C1 questions	Same as DSB QA

Table 1: Summary of the provided datasets for MT and QA.

the reference dataset we retrieve the 75 most similar Ukrainian sentences from the collected pool of translation data along with the associated sentence in CS or EN and aggregate them. We then deduplicate the aggregated set to retain only unique translation pairs. Overall, we get 321k and 251k CS–UK and, respectively, EN–UK translation pairs for training.

2.3 Data for Retrieval-augmented Generation

For the Ukrainian QA task, we employ retrieval-augmented generation (RAG) using pages from Wikipedia and 10 books on Ukrainian history, language, and literature (the same sources used by the winning team of the UNLP 2024 Shared Task (Boros et al., 2024)). We extract about 30k pages using the `wikipediaapi`[4] library, setting `max_depth = 2` to include relevant subcategories from the Ukrainian history, language, and literature categories.

3 Models and Algorithms

3.1 Qwen: The Underlying LLM

All our models are LoRA (Hu et al., 2021)-finetuned Qwen2.5-3B-Instruct models, and, thus, satisfy the shared task's parameter constraint. We use one model per language for all tasks.

3.2 Finetuning

Finetuning on MT and General QA Data For DSB and HSB, we combine the provided MT data, additional translations created by us, and translated MCQs to finetune Qwen2.5-3B-Instruct with LoRA (Hu et al., 2021) applied to all projection layers. For DSB/HSB, we use both the English and the translated version of each MCQ, in this format:

MCQ Prompt for DSB/HSB during training

```
en_context (if any)
dsb/hsb_context (if any)

Question: {en_question}
Question: {dsb/hsb_question}

Possible Answers: {en_possible_answers}
Possible Answers: {dsb/hsb_possible_ans}

Answer: {answer}
```

To extract the model's predicted answer for QA, we end the prompt with "`Answer:`" and compute the next-token probabilities for each option label. The answer is then taken as the label with the highest probability. Following Sanz-Guerrero et al. (2025), we evaluate the probabilities of the "␣X" tokens[5] (i.e., tokens formed by the preceding space *together* with the option label), as this approach has been shown to yield better performance and calibration.

In order to use this prompt at inference time, we need to translate the provided DSB/HSB questions into English. For, first we finetuned two separate translation models, one for DSB–EN and another for HSB–EN, using the same dataset we use for EN–HSB and EN–HSB model finetuning, but this time in the opposite direction. These two translation models are also based on Qwen2.5-3B-Instruct and trained with LoRA. During the development

[4] https://github.com/martin-majlis/Wikipedia-API

[5] Where "X" denotes one of the option labels.

phase, we observe that for both DSB and HSB QA, alphabetic option labels lead to better results than numeric labels due to label bias (Zheng et al., 2024). So, for training and evaluation, we use alphabetic option labels.

The prompt for MT is of the following format:

MT Prompt during Training

```
Translate this German sentence into Upper
Sorbian. Put it in this format: <hsb> {Upper
Sorbian translation} </hsb>
<de> {German Sentence} </de>
```

For Lower Sorbian, "Upper" is replaced by "Lower" and <hsb> and </hsb> are replaced by <dsb> and </dsb>.

For Ukrainian (UK), we train the model on MT and QA data described in Section 2.2.

We apply the default chat template for Qwen2.5 and train on complete instructions (system + user + assistant), as described in Shi et al. (2024). We use LoRA (Hu et al., 2021) for 10 epochs with an initial learning rate of $1e-4$ and a linear learning rate scheduler. We save the model checkpoint after every epoch. At the end of training, we select the model with the highest BLEU (Post, 2018; Papineni et al., 2002) score, i.e., we select the model using the MT development set.

3.3 Final Finetuning on MT and In-domain QA Data for DSB and HSB

After initial finetuning of the model, we conduct a second round of finetuning. This final model finetuning is also performed by applying LoRA to all projection layers. We follow the same procedure for DSB and HSB. The provided QA development sets for DSB and HSB contain a total of 158 questions each, from language difficulty levels A1-B2. In this second stage of finetuning, we use these 158 in-domain QA examples and the first 3k translation pairs used in the initial finetuning process. To mitigate data scarcity, we apply oversampling: each QA item is repeated five times. Then, we shuffle the MT and the oversampled QA set and finetune the first finetuned model again to improve domain alignment for QA. We add the 3k translation pairs to avoid catastrophic forgetting of the MT capability of the models. The final models are used for both MT and QA during evaluation. As our dataset for the second round of finetuning is small (approximately 3.75k MT and QA examples), we tune the learning rate, searching over $1e-4, 1e-5, 1e-6$, and $1e-7$. Four separate models are trained with these learning rates exactly for one epoch. In this learning rate searching process, we exclude the questions from B2 during finetuning. We select the best learning rate based on performance for both QA (56 questions of B2 level) and MT (the first 400 samples of the 4k dev set). We follow this approach for both languages. After this experiment, we chose $1e-4$ for DSB and $1e-6$ for HSB. Then, we finetune the initial models with these learning rates for two epochs on approximately 3.75k instructions, including the questions from B2, performing no validation.

3.4 Averaging Probabilities

During QA evaluation for DSB and HSB, we generate multiple responses for each question by shuffling the order of answer options. We perform this step to mitigate positional bias (Pezeshkpour and Hruschka, 2024) as much as possible. For questions with 2–3 options, we use all permutations, which are 2 and 6, respectively. For those with more than three options, we randomly sample 20 unique answer option orders. We compute the probability distribution over the answer options under the model for each order and average them; we select the option with the highest average likelihood as the final answer.

3.5 Retrieval-augmented Generation for Ukrainian QA

We segment the retrieved pages (see Section 2.3) into chunks of 512 characters with an overlap of 64 characters and embed them using Qwen2.5-3B-Instruct. For each chunk, mean pooling is applied over the token representations obtained from the last hidden states. Embeddings are stored in two separate `ChromaDB` indexes: one for history and another for language and literature. We make embedding of each question following the same way we apply for page chunks, mean pooling over all token representations. Using the subject indicated for each question, we conduct a search in the corresponding subject-specific index. At inference time, we retrieve the 5 most relevant chunks and use them as context alongside the question.

3.6 Few-shot In-context Learning for MT

For MT, we employ few-shot in-context learning using similarity-based retrieval, following Zebaze et al. (2025). For DSB/HSB, we embed each test-set German source sentence and retrieve the 5 most

Model	ChrF++	
	DSB	HSB
Qwen2.5-3B-Instruct + LoRA(S2)(P)	66.6	77.6
Qwen2.5-3B-Instruct + LoRA(S1)	67.5	77.5
Baseline	12.21	11.87

Table 2: ChrF++ scores for DSB and HSB. *LoRA(S1)* = one round of LoRA finetuning; *LoRA(S2)* = two rounds of LoRA finetuning. *P* indicates our primary submission.

similar sentences from the development set, along with their translations. For Ukrainian, we use Ukrainian sentences for embedding generation and retrieval.

4 Results and Discussion

MT for Lower and Upper Sorbian Table 2 shows that our proposed approach yields a significant improvement over the baseline. The baseline system achieves only 12.21 ChrF++ for DSB and 11.87 ChrF++ for HSB. The first round of LoRA finetuning, indicated by *S1*, already increases ChrF++ to 67.5 and 77.5 for DE–DSB and DE–HSB, respectively. The goal for our second round of finetuning (*S2*), is to adapt the model with in-domain QA data, while retaining the model's MT capability as much as possible. For HSB, *S2* slightly improves over S1 (77.6 vs. 77.5), but, for DSB, performance drops slightly, from 67.5 to 66.6.

Model	Accuracy (%)	
	DSB	HSB
Qwen2.5-3B-Instruct + NO-FT	54.3	57.1
Qwen2.5-3B-Instruct + LoRA(S1)	48.3	50.0
Qwen2.5 + LoRA(S1) Avg.	50.7	54.3
Qwen2.5-3B-Instruct + LoRA(S2)	48.3	50.5
Qwen2.5 + LoRA(S2) Avg. (P)	51.7	55.2
Baseline	45.85	42.86

Table 3: QA accuracy scores (A1–C1) for DSB and HSB. *LoRA(S1)* = one round of LoRA finetuning; *LoRA(S2)* = two rounds of LoRA finetuning; *Avg.* = average over multiple option orders; *NO-FT* = no finetuning, i.e., direct use of Qwen2.5-3B-Instruct. *P* indicates our primary submission.

QA for Lower and Upper Sorbian For DSB and HSB QA (Table 3), the baseline accuracies of 45.85% (DSB) and 42.86% (HSB) are surpassed by all finetuned variants. Interestingly, the non-finetuned Qwen2.5-3B-Instruct model outperforms the baseline substantially, particularly for HSB (+14.24 accuracy). However, LoRA finetuning (S1 and S2) slightly reduces overall accuracy compared to the non-finetuned model, likely due to the trade-off introduced by joint MT and QA finetuning. Averaging over the results for different answer option orders improves accuracy after both rounds of finetuning (S1 and S2). It helps more after the second round of finetuning, reaching 51.7% for DSB and 55.2% for HSB. The improvements over non-averaged results demonstrate that this straightforward method is effective for low-resource QA.

Model	ChrF++	
	CS–UK	EN–UK
Qwen2.5-3B-Instruct + LoRA	8.09	3.10
Baseline	3.48	0.40

Table 4: ChrF++ scores for CS–UK and EN–UK MT.

Model	Accuracy (%)
Qwen2.5-3B-Instruct + LoRA + RAG	35.82
Baseline	31.16

Table 5: Accuracy scores for Ukrainian QA. The shown model is our primary submission.

MT for Ukrainian The Ukrainian MT tasks (Table 4) are challenging due to the lack of training data provided by the shared task. The baseline ChrF++ scores of 3.48 (CS–UK) and 0.40 (EN–UK) reflect the difficulty level. By retrieving and curating translation data via similarity search and then finetuning a Qwen2.5-3B-Instruct model, our system achieves slight improvements over the baseline for both CS–UK (8.09) and EN–UK (3.10). Though performance increases, the improvement is not as big as for the DE–DSB/HSB MT tasks. A possible reason for this is a mismatch between the training, development, and test sets: we train our models on sentences, but the test set consists of large documents and lengthy conversations.

QA for Ukrainian For Ukrainian QA (Table 5), our proposed model, based on finetuning jointly on MT and QA in combination with RAG, improves over the baseline: 31.16% vs. 35.82%. This gain is smaller than for DSB and HSB QA, which we attribute to two factors: First, the QA dataset for Ukrainian requires some factual knowledge regard-

ing the Ukrainian language, history, and literature, which makes the task harder. Second, most LLMs are underexposed to the Cyrillic script, resulting in weaker tokenization, over-splitting of words, and a decreased quality of token representations (Boros et al., 2024). This results in poor-quality embeddings for Ukrainian sentences. As a result, retrieval quality degrades: semantically close passages are missed or under-ranked, and the injected context is less helpful.

5 Conclusion

In this paper, we present JGU Mainz's submission to the WMT25 Shared Task on LLMs with Limited Resources for Slavic Languages, addressing MT and QA in Ukrainian, Upper Sorbian, and Lower Sorbian. Our approach combines parameter-efficient finetuning of Qwen2.5-3B-Instruct with training data augmentation and RAG for Ukrainian QA. Our primary submissions outperform the provided baselines for all languages and tasks, achieving substantial ChrF++ gains for DE—DSB and DE-–HSB MT, as well as slight improvements for CS—UK and EN-–UK MT. For QA, averaging over order options increases accuracy for both DSB and HSB, while, for Ukrainian, we achieve moderate gains through RAG.

Acknowledgments

This work was supported by the Carl Zeiss Foundation through the TOPML and MAINCE projects (grant numbers P2021-02-014 and P2022-08-009I).

References

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: reasoning about physical commonsense in natural language. *CoRR*, abs/1911.11641.

Tiberiu Boros, Radu Chivereanu, Stefan Dumitrescu, and Octavian Purcaru. 2024. Fine-tuning and retrieval augmented generation for question answering using affordable large language models. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 75–82, Torino, Italia. ELRA and ICCL.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *CoRR*, abs/1905.10044.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

ELRC. 2022. ELRC Ukrainian Acts. `https://elrc-share.eu/repository/browse/eu-acts-in-ukrainian/71205868ae7011ec9c1a00155d026706d86232eb1bba43b691bdb6e8a8ec3ccf/`. [Online; accessed 05-August-2025].

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Matt Gardner Johannes Welbl, Nelson F. Liu. 2017. Crowdsourcing multiple choice science questions.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *arXiv:1910.11473v2*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

OPUS. 2003. OPS Website. `https://opus.nlpl.eu/`. [Online; accessed 05-August-2025].

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Nataliia Saichyshyna, Daniil Maksymenko, Oleksii Turuta, Andriy Yerokhin, Andrii Babii, and Olena Turuta. 2023. Extension Multi30K: Multimodal dataset for integrated vision and language research in Ukrainian. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 54–61, Dubrovnik, Croatia. Association for Computational Linguistics.

Mario Sanz-Guerrero, Minh Duc Bui, and Katharina von der Wense. 2025. Mind the gap: A closer look at tokenization for multiple-choice question answering with llms. *Preprint*, arXiv:2509.15020.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction tuning with loss over instructions. *Preprint*, arXiv:2405.14394.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge dataset and models for dialogue-based reading comprehension. *Preprint*, arXiv:1902.00164.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *CoRR*, abs/1811.00937.

Armel Randy Zebaze, Benoît Sagot, and Rachel Bawden. 2025. In-context example selection via similarity search improves low-resource machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1222–1252, Albuquerque, New Mexico. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

Krey-All WMT 2025 CreoleMT System Description: Language Agnostic Strategies for Low-Resource Translation

Ananya Ayasi
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA, USA
ananya.ayasi@gmail.com

Abstract

This work is a submission to the Creole MT Task conducted as a part of the Tenth Conference on Machine Translation (WMT '25) co-located with EMNLP 2025. The focus of the work was on the Seychellois Creole- English language pair by utilizing similar Creoles to make up for the scarcity of data in Creole language systems.

1 Introduction

The work of this paper is based on Robinson et al. (2024) which presented a one of a kind, cumulative dataset for Creole language MT. This is an important topic owing to the fact that many Creole language speakers live in areas where their language is in the minority and they could benefit from better machine translation systems.

The main objective behind this paper is to see if a certain language pair translation could benefit from similar languages when building an MT system. To that end, I picked Seychellois Creole (crs)- English (eng) pair and to supplement the model training Mauritian Creole (mfe), French Guianese Creole (gcr), Louisiana Creole (lou) and Réunion Creole (rcf) were chosen. The linguistic proximity of these Creoles were studied in Papen (1978). Many Creoles also have linguistic relationships with high resource languages which gives better prospects for cross-lingual transfer (Lent et al., 2024) and hopefully can be studied in a future iteration of this paper.

In addition to the shared-task submission for crs–eng, this work evaluates the broader applicability and limits of unified tagging by comparing performance across multiple French-lexifier Creoles as well as Tok Pisin, a typologically distant creole, thereby situating the approach within a linguistically grounded framework rather than a purely engineering exercise.

2 Data

Since this work was part of the constrained track, the data and base model used were provided by Robinson et al. (2024). As mentioned above, the 5 Creole languages used were chosen based on their linguistic proximity.

Seychellois Creole and Mauritian Creole are widely regarded as the closest pair among the varieties considered here, a connection that stems largely from historical migration patterns (Baker et al., 1982). Seychelles was primarily settled from Mauritius in the late 1700s, and while there are lexical and phonological differences, several scholars have noted that they may be treated as regional variants of a single linguistic system rather than entirely separate languages (Michaelis and Rosalie, 2013) (Ramsurrun et al., 2024).

Translation Pair	Train	Val	Test
crs-eng	2,070	107	222
mfe-eng	19,100	472	811
gcr-eng	60	28	38
lou-eng	1,570	92	174
rcf-eng	191	34	28

Table 1: Number of training, validation, and test sentence pairs for each Creole–English translation dataset used in this study.

Réunion Creole, while geographically proximate, has a distinct and earlier colonization history, resulting in greater grammatical divergence from Mauritian. Nevertheless, commonalities in the pronominal system, such as the use of *zot* for both second and third-person plural, and shared plural markers like *ban(de)*, point to the effects of geographical continuity and inter-island contact. Lexical similarities, including borrowings from Malagasy and Indo-Portuguese, reinforce the role of shared areal features despite significant typological differences (Papen, 1978).

Expanding beyond the Indian Ocean, Louisiana Creole and French Guianese Creole also exhibit no-

Proceedings of the Tenth Conference on Machine Translation, pages 1158–1165

table structural parallels with Mauritian, underscoring broader transoceanic affinities among French-lexifier Creoles. (Pfänder, 2013) (Klingler and Neumann-Holzschuh, 2013)

For MT research, this interconnected yet varied typological landscape offers opportunities for cross-lingual transfer learning, provided that models are designed to exploit shared structures without conflating distinct grammatical systems.(Grant and Guillemin, 2012)

As far as Creoles go, I have also chosen Tok Pisin as a linguistically dissimilar language to Seychellois Creole to demonstrate how this strategy is effective only when the varieties share substantial structural overlap.

3 Methodology

This work adopts a language-agnostic machine translation strategy for low-resource Creole–English translation, building on multilingual pretraining and fine-tuning techniques that have been shown to benefit typologically related languages in low-data regimes (Johnson et al., 2017); (Aharoni et al., 2019); (Conneau et al., 2020).

I call the approach "language-agnostic" because the model is not given any explicit linguistic rules or handcrafted features about Creole varieties; instead, all source languages are treated under a unified tag, allowing the system to learn transfer purely from shared representations without relying on language-specific annotations.

3.1 Base Model and Setup

The base model is kreyol-mt-pubtrain based on mBART (Namdarzadeh et al., 2023) as distributed in the constrained track by Robinson et al. (2024). This model is pretrained on a variety of languages, including several French-based Creoles, providing a strong initialization for transfer learning. All training experiments use the HuggingFace Transformers library (Wolf et al., 2020) with mixed precision and early stopping where indicated.

3.2 Tagging Strategies

The main focus of this paper was to see if other similar languages can be used to enhance the MT system for a certain language pair or rather language-agnostic machine translation as inspired by (Chen and Zhang, 2024). I mainly experimented with two techniques for the same, "All Kreyols" and "Specialized".

Following Johnson et al. (2017), all input sequences were prepended with a target language tag. In the "All Kreyols" condition which was the language-agnostic approach, all Creole data received the <2crs> tag regardless of source variety, treating all Creoles as dialectal variants for the purpose of model conditioning.

"Partial Kreyall" was basically a variant of "All Kreyoles" where even though all the languages used the same tag, only partial fine-tuning was done. In the "Specialized" condition which was language-specific, each Creole variety retained its own target tag (e.g., <2mfe>, <2lou>), enabling the model to distinguish among them. In this case, I experimented with both full fine-tuning and partial fine-tuning.

3.3 Data Augmentation via Upsampling

It is to be noted that in Table 1 mfe-eng language pairs in the train set was almost twice as that of crs-eng. Given the extreme imbalance in dataset sizes, upsampling of the target language data was employed to avoid underrepresentation of Seychellois Creole. Without intervention, batches are dominated by mfe examples, which can lead to underrepresentation of crs and limiting domain adaptation.

Inspired by Sennrich et al. (2016) and Robinson et al. (2022), initial experiments used a 5x upsampling factor for crs-eng. This ensures that crs examples are seen as frequently as those from the largest dataset, allowing the model to retain multilingual benefits while focusing on the target low-resource language. Later, a 10x factor was tested, especially in the "Specialized" setting, to further balance the distribution. However, it did not exactly lead to better results as will be discussed later.

3.4 Fine-tuning Strategies

Two broad strategies were mainly tested for fine-tuning. The first one being without any freezing, thereby fine-tuning all the parameters of the given kreyol-mt-pubtrain model.

The second one being partial freezing where only the last 4 to 6 encoder layers, all decoder layers, and optionally shared embeddings were frozen. This follows the intuition from Kirkpatrick et al. (2017) and Pfeiffer et al. (2021) that retaining most pretrained parameters can preserve generalization while adapting only the task-relevant parts.

Additional architectural regularization was explored by modifying dropout rates for feedforward, attention, and activation layers. In the notation

dropout=0.2/0.05/0.05 in Table 4, the three values correspond respectively to the model's feedforward dropout, attention dropout, and activation dropout. For instance, in kreyol-mt-pubtrain's configuration, these would be set via:

```
model.config.dropout = 0.2
# feedforward layer dropout
model.config.attention_dropout = 0.05
# attention mechanism dropout
model.config.activation_dropout = 0.05
# activation function dropout
```

These rates control the probability of randomly zeroing out elements during training to prevent overfitting. Adjusting them independently allows for targeted regularization. Higher dropout in feedforward layers can reduce co-adaptation in dense transformations, while lower dropout in attention and activation layers can help preserve sequence modeling capacity without severely impacting convergence.

3.5 Optimization

All models were optimized using AdamW, which decouples weight decay from the gradient-based updates, preventing it from accumulating in the momentum or variance terms. Weight decay values ranged from 0.1 to 0.3, and the learning rate was fixed at 1E-5 to mitigate catastrophic forgetting in the multilingual setting. Warmup schedules of 500–1000 steps were applied, followed by linear, cosine, or constant learning rate decay. In certain configurations, label smoothing of 0.1–0.2 (Szegedy et al., 2015) was incorporated to improve generalization in the low-resource MT setting. All hyperparameters were chosen based on prior work in low-resource MT and small-scale tuning on the validation set.

4 Experiments

The experiments compare baseline, All Kreyols, Partial Krey-all, and Specialized setups. Performance was evaluated using BLEU (Papineni et al., 2002) and chrF (Popović, 2015) on the held-out Seychellois Creole–English test set.

4.1 Language Embeddings

In order to truly understand and corroborate prior linguistic research pertaining to the proximity of the French-lexifier Creoles, kreyol-mt-pubtrain's encoder embeddings for them and Tok Pisin (tpi) were visualized by sampling 60 sentence pairs from each dataset and converting their source sentences into 1024-dimensional vectors using the encoder.

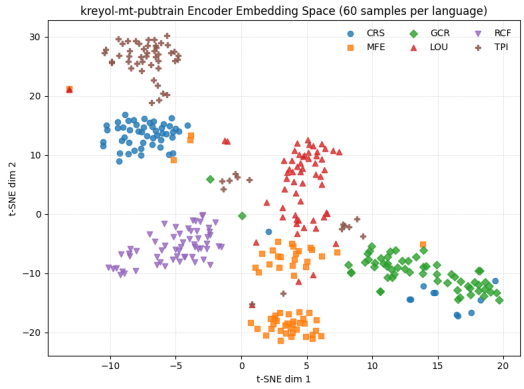


Figure 1: t-SNE visualization of kreyol-mt-pubtrain encoder embeddings for 60 sampled sentences from each language (crs = Seychellois Creole, mfe = Mauritian Creole, gcr = French Guianese Creole, lou = Louisiana Creole, rcf = Réunion Creole, tpi = Tok Pisin). The five French-lexifier Creoles form distinct yet proximate clusters, reflecting their structural similarity, whereas Tok Pisin is clearly separated in embedding space, highlighting its typological distance.

These high-dimensional vectors were then reduced to two dimensions via t-SNE (Cai and Ma, 2022), which preserves local similarity structure while making patterns viewable in a scatter plot. The resulting "t-SNE dim 1" and "t-SNE dim 2" axes have no direct linguistic meaning. Their role is to position points so that those close in the plot were also close in the original embedding space. Each language was assigned a unique color and marker, revealing how kreyol-mt-pubtrain clusters them based on learned representations.

The plot shows that the five French-lexifier Creoles form distinct but proximate clusters, with some overlap among related varieties such as gcr, mfe and lou. crs is positioned near rcf and partially overlaps with gcr, aligning with the shared French-derived vocabulary and grammar across these languages. By contrast, tpi is kept well apart from these clusters, reflecting its typological distance from the French-lexifier group. This separation reinforces the observation that unified tagging works best for closely related languages, where embeddings occupy overlapping regions and facilitate cross-lingual transfer (Ponti et al., 2021).

4.2 Baseline

The baseline fine-tuned the given kreyol-mt-pubtrain with no freezing and <2crs> tag for all data. The best run achieved BLEU = 34.61, chrF

= 60.84 after 15 epochs with a linear scheduler and 500 warmup steps. To ensure comparability, this baseline was reproduced under the same constrained-track conditions, and the scores are in line with those reported in the shared-task paper, confirming that my setup is consistent with the official baseline.

4.3 All Kreyols/ Krey-All

Here, all Creole varieties were merged under the <2crs> tag. The best configuration reached BLEU = 35.44, chrF = 61.71 using 20 epochs (stopping at epoch 18), cosine scheduling, LR = 1E-5, warmup = 1000, and label smoothing = 0.1. This shows gains over the baseline, suggesting that cross-variety pooling helps despite grammatical differences.

4.4 Partial Krey-all

For this setting, partial freezing was applied, testing different numbers of unfrozen encoder layers, decoder layers, and embedding sharing. The results were mixed with some configurations having degraded performance (e.g., BLEU = 20.14 for unfreezing last 2 encoder/decoder layers), while others recovered to near-baseline levels. The gains were limited, indicating that aggressive layer freezing may not suit closely related low-resource varieties.

4.5 Specialized

This setup retained distinct tags per variety, with and without crs upsampling. The highest performance here (BLEU = 35.19, chrF = 61.14) was achieved by disabling early stopping, allowing the model to converge fully over 20 epochs. Notably, 10x crs upsampling did not yield improvements over the baseline or the settings with 5x upsampling. The findings suggest a saturation effect, where oversampling beyond 5x no longer improves performance because the model's representation of crs is already well-established.

4.6 Other Language Pairs

To further validate the generality of the Krey-All unified tagging strategy, I extended experiments beyond the official shared-task pair (crs–eng) to additional Creole–English directions. For the French-lexifier creoles (crs, mfe, rcf, lou, gcr), I trained each pair by combining its own data with data from the other French-lexifier creoles, while preserving their respective language tags. This ensured that the unified strategy was tested symmetrically across all related varieties. The results, shown in Table 3, indicate consistent improvements across all of these creoles, with gains ranging from +0.9 to +1.3 BLEU.

To further stress-test the approach, I then applied the same unified tagging setup to Tok Pisin (tpi) by pooling it together with all the French-lexifier creoles. Unlike the others, tpi is typologically distant, and the results confirm that the strategy does not generalize well in this case (–1.28 BLEU). Importantly, to maintain the sanctity of the experiments, tpi data was not used when training the French-lexifier systems. This separation highlights the core finding: unified tagging provides consistent benefits when applied to closely related creoles, but degrades performance when applied to a structurally different language.

All experiments were conducted on an NVIDIA A100 GPU. Also note that for the shared task submission, crs–eng Krey-All was submitted as the primary system, while crs–eng Partial Krey-All and crs–eng Specialized Krey-All were submitted as contrastive systems 1 and 2, respectively.

5 Results and Analysis

The language-agnostic tagging strategy which was the main goal of this paper proved most effective, as shown by two clear outcomes. First, tagging crs across all Creoles consistently yielded higher scores compared to runs where tags were language-specific. Second, even when crs data was oversampled by 10x, scores did not improve over the 5x runs, despite mfe data being more abundant, indicating that mfe served as a strong substitute for crs in the shared tag setup. This aligns with prior findings that shared tags encourage parameter sharing across related low-resource languages (Johnson et al., 2017); (Sachan and Neubig, 2018).

No freezing continued to outperform partial freezing, suggesting that kreyol-mt-pubtrain's multilingual pretraining already aligns Creole varieties in the embedding space. Scheduler and label smoothing choices impacted results more than weight decay, with cosine or linear schedules and smoothing = 0.1 performing best overall.

6 Conclusion

This study examined a unified tagging strategy using kreyol-mt-pubtrain to improve translation quality for the low-resource Seychellois Cre-

Table 2: Best results in each category for Creole–English MT.

Technique	Details	BLEU	chrF
Baseline	No freezing	34.61	60.84
Partial Krey-all	Unfreeze last 4 encoder / all decoder layers+ shared embeddings, dropout=0.2/0.05/0.05	34.72	61.11
Specialized	No freezing and no early stopping	35.19	61.14
All Kreyols	No freezing	**35.44**	**61.71**

Creole–Eng	Baseline	Krey-All
crs–eng	34.61	35.44
rcf–eng	41.00	42.12
lou–eng	32.72	33.65
gcr–eng	49.89	51.21
mfe–eng	25.23	26.29
tpi–eng	53.25	51.97

Table 3: **BLEU scores** for baseline vs. Krey-All unified tagging across Creole–English pairs. French-lexifier creoles consistently improve under unified tagging, while Tok Pisin (tpi), a typologically distant language, shows degraded performance.

ole–English pair. By tagging crs across all Creole varieties, the model achieved higher scores than language-specific tagging. The absence of gains from 10x crs upsampling, despite the availability of larger mfe data, further indicated that Mauritian Creole served as an effective proxy for crs in the shared-tag setup. These findings reinforce that related Creole languages can transfer knowledge efficiently when trained under a unified representation.

While absolute gains are modest (<1 BLEU/chrF), they are consistent across runs and show that leveraging related Creoles can stabilize low-resource training. Moreover, since the constrained track provides the same pretraining data for both base and fine-tuning, the improvements are naturally limited compared to unconstrained settings.

Beyond crs–eng, applying the same strategy to other Creole–English directions (rcf, lou, gcr,mfe) produced similar improvements, confirming that the benefits of unified tagging generalize across French-lexifier Creoles. In contrast, Tok Pisin (tpi), a typologically distant creole, did not benefit, which corroborates the claim that this approach is most effective when applied to closely related languages.

No-freezing fine-tuning proved most effective, suggesting that kreyol-mt-pubtrain's multilingual pretraining already aligns Creole varieties well in embedding space. The best system, trained on all Creoles without freezing, reached a BLEU of 35.44 and chrF of 61.71 for crs-eng, outperforming the baseline. Overall, cross-lingual transfer with linguistically related low-resource languages emerges as a promising strategy in constrained MT settings, particularly when paired with balanced exposure and a unified tagging scheme.

If there is one takeaway from this paper, it is that prior linguistic research remains vital even in the age of large language models. More often than not, leveraging linguistic insight can be just as critical as scaling data volume. (Gu, 2025)

Limitations

While the results are promising, there are several limitations to this work:

- Data imbalance: The extreme disparity in dataset sizes (e.g., 2K pairs for crs-eng vs. 19K for mfe-eng) may cause the model to underfit the smallest dataset even after upsampling. It is also to be noted that in the original paper, the scores for mfe-eng were much lower than crs-eng despite the larger training set. This should be further investigated in upcoming iterations.
- Domain mismatch: Training corpora for different Creoles may differ in domain, register, or orthography, potentially introducing noise (Karakasidis et al., 2023).
- Model constraints: We were limited to the kreyol-mt-pubtrain model in the constrained track, restricting architectural or pre-training modifications.
- Evaluation scope: BLEU and chrF scores, while informative, do not capture deeper semantic adequacy or fluency. No human evaluation was performed.

Ethics Statement

This work uses only publicly released datasets from Robinson et al. (2024) and adheres to the constraints of the WMT 2025 Creole MT Task. No personally identifiable information (PII) is present in the training or evaluation data.

While the goal is to improve accessibility for speakers of low-resource Creole languages, I acknowledge that MT systems may propagate biases present in the source data, especially given the historical and sociolinguistic contexts of Creole-speaking communities. Deployment of such systems should therefore be accompanied by careful evaluation in real-world contexts, with community feedback guiding improvements.

Additionally, these models are intended to augment human communication rather than replace professional translators, especially in sensitive domains such as legal, medical, or governmental communication.

Acknowledgements

The author thanks Nate (author of Robinson et al. (2024)) for his invaluable guidance and mentorship throughout this work. Gratitude is also extended to family (Seema, Ajai, Anwaya) and friends (Ishwrya, Balaji, Maanasa, Sruthi) for their unwavering support during the course of this research. This work is dedicated to Simi, Eldin, Ewan, and Ryan.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

Baker, Philip; Corne, and Chris. 1982. *Isle de France Creole: Affinities and origins*. Karoma, Ann Arbor, MI.

T. Tony Cai and Rong Ma. 2022. Theoretical foundations of t-sne for visualizing high-dimensional clustered data.

Xiao Chen and Chirui Zhang. 2024. Language-agnostic zero-shot machine translation with language-specific modeling. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Anthony Grant and Diana Guillemin. 2012. The complex of creole typological features: The case of mauritian creole. *Journal of Pidgin and Creole Languages*, 27(1):48–104.

Wenshi Gu. 2025. Linguistically informed chatgpt prompts to enhance japanese-chinese machine translation: A case study on attributive clauses. *PloS one*, 20(1):e0313264.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Georgios Karakasidis, Nathaniel Robinson, Yaroslav Getman, Atieno Ogayo, Ragheb Al-Ghezi, Ananya Ayasi, Shinji Watanabe, David R Mortensen, and Mikko Kurimo. 2023. Multilingual tts accent impressions for accented asr. In *International Conference on Text, Speech, and Dialogue*, pages 317–327. Springer.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Thomas A. Klingler and Ingrid Neumann-Holzschuh. 2013. Louisiana creole. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *The survey of pidgin and creole languages. In "The survey of pidgin and creole languages". Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford University Press, Oxford.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, Anders Søgaard, and Johannes Bjerva. 2024. Creoleval: Multilingual multitask benchmarks for creoles.

Susanne Maria Michaelis and Marcel Rosalie. 2013. Seychelles creole. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus

Huber, editors, *The survey of pidgin and creole languages. In "The survey of pidgin and creole languages". Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford University Press, Oxford.

Behnoosh Namdarzadeh, Sadaf Mohseni, Lichao Zhu, Guillaume Wisniewski, and Nicolas Ballier. 2023. Fine-tuning MBART-50 with French and Farsi data to improve the translation of Farsi dislocations into English and French. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 152–161, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Robert Papen. 1978. *The French-based Creoles of the Indian Ocean: An analysis and comparison*. University of California, San Diego. Dissertation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning.

Stefan Pfänder. 2013. Guyanais. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *The survey of pidgin and creole languages. In "The survey of pidgin and creole languages". Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford University Press, Oxford.

Edoardo Maria Ponti, Julia Kreutzer, Ivan Vulić, and Siva Reddy. 2021. Modelling latent translations for cross-lingual transfer.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Neha Ramsurrun, Rolando Coto-Solano, and Michael Gonzalez. 2024. Parsing for mauritian creole using universal dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12622–12632.

Nathaniel Robinson, Perez Ogayo, Swetha Gangu, David R. Mortensen, and Shinji Watanabe. 2022. When is tts augmentation through a pivot language useful?

Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A. Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Dean Stutzman, Bismarck Bamfo Odoom, Sanjeev Khudanpur, Stephen D. Richardson, and Kenton Murray. 2024. Kreyòl-mt: Building mt for latin american, caribbean and colonial african creole languages.

Devendra Sachan and Graham Neubig. 2018. Parameter sharing methods for multilingual self-attentional translation models. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 261–271, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

A Additional Results

Table 4: Full experimental results (BLEU/chrF on crs-ENG).

Technique	Details	Epochs	Stopped at	LR Sched.	Warmup	BLEU	chrF
Baseline	No freezing	15	15	linear	500	**34.61**	**60.84**
All Kreyols	No freezing	5	5	linear	750	34.88	61.26
All Kreyols	No freezing	20	11	linear	500	34.43	61.18
All Kreyols	No freezing	20	18	cosine	1000	**35.44**	**61.71**
All Kreyols	No freezing	30	14	linear	750	35.06	61.56
All Kreyols	No freezing	25	10	cosine	1000	35.25	61.49
All Kreyols	No freezing	20	13	constant	1000	35.11	61.45
All Kreyols	No freezing	30	4	cosine	1000	34.70	60.95
Partial Krey-all	Unfreeze last 2 encoder/decoder layers	5	4	cosine	1000	20.14	48.55
Partial Krey-all	Unfreeze last 2 encoder/decoder layers + shared embeddings	5	5	cosine	1000	33.90	60.24
Partial Krey-all	Unfreeze last 4 encoder / all decoder layers + shared embeddings	20	7	cosine	1000	33.91	60.71
Partial Krey-all	Unfreeze last 4 encoder / all decoder layers + shared embeddings; dropout=0.3/0.1/0.1	20	7	cosine	1000	33.91	60.71
Partial Krey-all	Unfreeze last 4 encoder / all decoder layers + shared embeddings, dropout=0.2/0.05/0.05	20	15	cosine	1000	**34.72**	61.11
Partial Krey-all	Unfreeze last 6 encoder / all decoder layers + shared embeddings, no dropout	20	10	cosine	1000	34.20	**61.15**
Specialized	Unfreeze last 4 encoder / all decoder layers + shared embeddings, dropout=0.2/0.05/0.05	20	7	cosine	1000	34.51	60.27
Specialized	Unfreeze last 6 encoder / all decoder layers + shared embeddings, no dropout	20	13	cosine	1000	34.09	60.26
Specialized	No freezing	20	5	cosine	1000	34.81	60.93
Specialized	No freezing + crs upsampled 10×	5	5	cosine	1000	34.66	60.56
Specialized	No freezing and no early stopping	20	20	cosine	1000	**35.19**	**61.14**

EdinHelsOW WMT 2025 CreoleMT System Description: Improving Lusophone Creole Translation through Data Augmentation, Model Merging and LLM Post-editing

Jacqueline Rowe[1], Ona de Gibert[2], Mateusz Klimaszewski[3], Coleman Haley[1], Alexandra Birch[1], Yves Scherrer[4]
[1]University of Edinburgh [2]University of Helsinki
[3]Warsaw University of Technology [4]University of Oslo
Correspondence: jacqueline.rowe@ed.ac.uk

Abstract

In this work, we present our submissions to the unconstrained track of the System subtask of the WMT 2025 Creole Language Translation Shared Task. Of the 52 Creole languages included in the task, we focus on translation between English and seven Lusophone Creoles. Our approach leverages known strategies for low-resource machine translation, including back-translation and distillation of data, fine-tuning pre-trained multilingual models, and post-editing with large language models and lexicons. We also demonstrate that adding high-quality parallel Portuguese data in training, initialising Creole embeddings with Portuguese embedding weights, and strategically merging best checkpoints of different fine-tuned models all produce considerable gains in performance in certain translation directions. Our best models outperform the baselines on the Task test set for eight out of fourteen translation directions. When evaluated on decontaminated test sets, they surpass the baselines in all directions.

1 Introduction

The introduction of the first Shared Task for Creole language machine translation (MT) (Robinson et al., 2025) is emblematic of the increased attention that Creole languages have received in the field of Natural Language Processing in recent years, both as individual languages (Robinson et al., 2022; Dabre et al., 2014; Lent et al., 2021; Dabre and Sukhoo, 2022; Rowe et al., 2025) and in multilingual modeling efforts (Robinson et al., 2024; Lent et al., 2024). Building on the latter, this Shared Task covers over 50 Creole languages from a range of geographical and linguistic contexts. Some are relatively high-resourced; for example, Haitian Creole and Papiamento are supported in Google Translate and many others are institutionalised as official or educational languages (Robinson et al., 2024). Others are extremely low-resource languages or even critically endangered or extinct.

The Shared Task invites submissions of data and systems serving MT between any of the Creole languages and either English or French, with the existing Kreyòl-MT (Robinson et al., 2024) and CreoleVal (Lent et al., 2024) translation models serving as baselines. In this submission, we develop systems to translate between English (eng) and seven Lusophone[1] Creoles: Angolar (aoa), Annobonese (fab), Guinea-Bissau Creole (pov), Kabuverdianu (kea), Papiamento (pap), Principense (pre) and Sãotomense (cri).[2] This set includes relatively high-resource Creoles (like pap and kea) and extremely low-resource ones (like aoa, fab and pre).

In our submission, we utilise known strategies for low-resource MT as well as techniques designed to leverage the linguistic relationship between our seven Creoles of focus and Portuguese (por). In particular, we contribute the following:

- We collate additional parallel and monolingual data for pap, pov, kea and cri (Sections 3.1.2 and 3.1.3).
- We augment the training data with high-quality parallel eng-por data, synthetic parallel data created via back-translation, and distilled data created via forward-translation (Section 3.2).
- We fine-tune three pretrained multilingual base models with different combinations of data and initialisation strategies (Section 4.2).
- We apply model merging to further improve translation performance (Section 4.3).
- We post-edit system outputs using LLMs and bilingual lexicons, improving performance for five translation directions (Section 4.4).

We release our code in our Github repository.[3]

[1]Creoles which are related to Portuguese.

[2]We focus on translation between these seven Creoles and English due to availability of test data, but future work could expand to Creole–Portuguese translation.

[3]https://github.com/JacquelineRowe/EdinHelsOW_CreolesMT. Due to the copyright terms of most of our data sources, we do not publicly share our dataset. It is available

Proceedings of the Tenth Conference on Machine Translation, pages 1166–1182
November 8-9, 2025

2 Related Work

Robinson et al. (2024) release four versions of Kreyòl-MT **(KMT)**, a translation model which supports all seven of our Creole languages of focus. The four versions are created by training on both public and private datasets, and training both from-scratch and fine-tuning an existing model. For fine-tuning, they use `many-to-many (m2m)` **mBART-50** (Tang et al., 2021), a multilingual version of mBART (Liu et al., 2020) fine-tuned for translation between 50 languages, as the base model. mBART is a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages using the BART objective (Lewis et al., 2020). Lent et al. (2024) also fine-tune `m2m` **mBART-50** on a different set of Creole languages including pap.

While the models released in Robinson et al. (2024) and Lent et al. (2024) are the strongest baselines for MT for Creoles in general, some of our seven languages of focus are also included in other prior work on MT. The No Language Left Behind (**NLLB**) translation model excels at translation of low-resource languages, and supports pap and kea (as well as three other Creoles not included in our study) (NLLB Team et al., 2022). The training data curated as part of the NLLB effort include less than 10 bitexts for each Creole, but 28M monolingual sentences in pap and 300k in kea. The FLORES-200 evaluation dataset was also translated into both Creoles in this context.

Both kea and pap are featured in **PanLex**,[4] a massive, open-access online lexicon covering over 5,000 languages (Kamholz et al., 2014); but only pap is supported in **GATITOS**, a smaller, higher-quality parallel lexicon for low-resource languages developed by Jones et al. (2023). These lexical resources have been used to improve low-resource MT performance for Creoles. Following prior work using LLMs to post-edit machine translation system outputs to correct errors (Xu et al., 2024; Chen et al., 2024; Hus et al., 2025), Nielsen et al. (2025) showed that including the entire GATITOS lexicon in such post-editing prompts can improve ChrF scores and reduce lexical confusion, including for pap-eng MT. Similarly, Hus and Anastasopoulos (2024) showed improvements of over 15 ChrF++ in eng-kea MT by post-editing using an LLM with prompts including parallel words and sentences extracted from the kea PanLex dataset.

The question of how training data from related languages can improve MT for Creoles remains open (Lent et al., 2022). Ma et al. (2025) showed that the speech foundation model Whisper (Radford et al., 2023) performs surprisingly well on kea-eng speech translation (despite having not been trained on kea speech) when the por language code is used for decoding, which they hypothesise is due to pronunciation similarities between the two languages. Conversely, Fekete et al. (2025) demonstrated that parameter efficient fine-tuning via language adapters improves MT for three Creoles (including pap) regardless of whether the adapters were trained on related languages, unrelated languages, or even random noise, indicating that language adapters improve performance due to regularization rather than cross-lingual transfer.

3 Data

In this section, we briefly describe the data provided by the task organisers and the additional data we collect and create for model training. Our novel data sources are documented in full in Table 6 in Section A.

3.1 Data Collection

3.1.1 Organiser-Provided Data

To train their models, Robinson et al. (2024) gathered data for 43 Creoles from multilingual datasets, extracting parallel and monolingual texts from websites, Wikipedia collections, educational materials, religious texts and other sources where available. Some of their data remains private due to copyright reasons, but their public training and development splits ($\text{Train}_{\text{KMT}}$ and Val_{KMT}) form the official training data for the Shared Task. Robinson et al. (2024) also provide a public test split (Test_{KMT}), which we do not use as training data.[5]

For our seven Creoles of focus, the publicly available resources parallel with eng from Robinson et al. (2024) vary in size and domain. The datasets for pov, pre, aoa, cri and fab have between 170 and 450 parallel aligned sentences from

to academic researchers for non-commercial purposes upon request; please contact the lead author for license agreement and access.

[4] At the time we conducted our study, PanLex was not accessible online and so we did not use this resource for kea in our work.

[5] While we did not use Test_{KMT} data to train our models, we did evaluate our models' performance on this public test split in order to make modelling decisions, prior to the announcement that the official Shared Task test set would be identical to Test_{KMT}.

educational materials, collected from the APiCS corpus (Michaelis et al., 2013). In contrast, the parallel datasets for kea and pap are larger and more diverse, both drawing data from FLORES-200 dev and NLLB train (NLLB Team et al., 2022) as well as APiCS (Michaelis et al., 2013). The public pap dataset[6] also includes bitexts from the Online library of The Church of Jesus Christ of Latter-day Saints[7], LegoMT (Yuan et al., 2023), Tatoeba[8], and Wikipedia, as well as a bilingual lexicon.[9] Parallel sentences with languages other than eng are available for pap and kea, but we include only parallel data with eng.

3.1.2 Additional Parallel Data

To augment the official task data, we collect additional data parallel with eng for pap, pov and kea.[10] As is common with low-resource languages, much of the publicly-available parallel data sources we could find for each language are religious in nature (Siddhant et al., 2022). We collect aligned Bible verses (pap and pov) and aligned sentences from available editions of Jehovah's Witnesses Watchtower (JWW) series[11] (pap, pov and kea). We also collect non-religious parallel sentences from a random sentence generator (pap), an article about internet access (pov), and the glosses from a por-pov bilingual dictionary (we translate the por glosses into eng using Google Translate).

Portuguese Since our focus is on Lusophone Creoles, we hypothesise that adding high-quality eng-por data can improve transfer learning. We download the eng-por Tatoeba Translation Challenge Dataset (Tiedemann, 2020), which is a collection of all data in OPUS, shuffled and deduplicated. We use the corresponding Bicleaner-AI (Zaragoza-Bernabeu et al., 2022) scores[12] to aggressively filter the dataset. Bicleaner-AI is a neural metric that estimates how likely it is that a sentence pair is a translation. We keep only sentence pairs with a Bicleaner-AI score of 1.0 to ensure high quality, leaving us with a seed dataset of 112k sentences (representing 0.03% of the total Tatoeba dataset).

3.1.3 Additional Monolingual Data

We also collect monolingual Creole data, including a high school textbook (kea), a blog series (kea), glosses from an unpublished monolingual dictionary (pov) and transcriptions of a documentary (pov). The JWW Series (see Section 3.1.2) in cri is hosted on a different website from the eng, pap, pov and kea versions; as this makes it impossible to align the cri data with the eng data, we instead collect JWW as a monolingual resource for cri.

3.1.4 Lexicons

In order to experiment with post-editing with LLMs and lexicons, as demonstrated in Nielsen et al. (2025), we collect bilingual lexicons for each of our seven Creoles of focus. For aoa, we could not find a publicly-available lexicon, and instead manually curate a small set of parallel lexical items using word-aligned entries from IMT Vault.[13] For pap, we use both the GATITOS lexicon (Jones et al., 2023) and a newly collected traditional lexicon.

3.2 Synthetic Data

We backtranslate all sources of monolingual data into eng using the KMT model that scores the highest ChrF on the KMT test set for that language pair.[14] We also use `kreyol-mt` (the single best KMT model) as a 'teacher' model, using it to forward translate the eng sentences from the pap, kea, pov and cri parallel datasets into each Creole via Sequence-Level Distillation (Seq-KD) (Kim and Rush, 2016).[15] These distilled datasets allow us to train models which better imitate the distribution output of the KMT model at sentence-level.

[6] The private pap-eng dataset (used for model training but not publicly released) includes additional parallel data from CreoleVal (Lent et al., 2024), a textbook, the JHU bible corpus (McCarthy et al., 2020), the QED corpus (Lamm et al., 2021) and Ubuntu texts from the OPUS corpus.

[7] https://www.churchofjesuschrist.org/study?lang=pap. This dataset was shared directly by the organisers as it is not on HuggingFace yet.

[8] https://tatoeba.org/en/downloads

[9] https://www.scribd.com/document/119363393/Parleremo-English-Papiamento-Papiamento-English-Dictionary-1ed

[10] We later found small parallel resources for aoa, fab and pre; while it was too late to include these sources in our model training, we list these sources in Table 6 for future reference.

[11] JWW is a monthly Bible study resource which is mostly about religious matters but also includes some discussion of more general topics.

[12] https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/d34a89ac102fd236503a1911dd1050564bf4e682/BicleanerScores.md

[13] https://imtvault.org/?languageiso6393%5B0%5D=aoa

[14] `kreyol-mt` for cri and kea; `kreyol-mt-pubtrain` for pap and pre; and `kreyol-mt-scratch` for pov. We used the publicly available Test_{KMT} set to select which models to use for back-translation before realising that the Shared Task test set would be identical to the publicly available test set.

[15] We do not use distillation for aoa, fab and pre because the KMT model demonstrates ChrF scores which are too low to generate reasonable forward translations.

3.3 Data Pre-processing

We use all novel collected data for training and evaluation, except the bilingual lexicons which we reserve for post-editing experiments. We first remove any pairs of parallel sentences from our novel datasets where either the source (`src`) or target (`tgt`) sentence is in that language pair's Test_{KMT} dataset, to ensure we do not train on any test data. We then split out 10% of our novel `cri`, `kea`, `pap` and `pov` data (up to a limit of 1,000 sentences) for both validation and test data. We combine our training and validation splits with $\text{Train}_{\text{KMT}}$ and Val_{KMT} respectively, but keep Test_{KMT} separate from our own test data ($\text{Test}_{\text{Ours}}$) for evaluation purposes.

To clean each data split, we remove duplicates, empty or identical `src`/`tgt` pairs, and pairs where `src` or `tgt` have more than 150 or fewer than three words. We also discard pairs where the ratio of the length of the `src` to the `tgt` sentence is unusually high or low, following Robinson et al. (2024). Finally, we normalise special characters like quotes, dashes, and Unicode Hex codes.

We noted that several sentences[16] from $\text{Train}_{\text{KMT}}$ and $\text{Validation}_{\text{KMT}}$ included multiple `eng` glosses for a single Creole sentence. For example, the `cri` sentence "*Ê tava ka vivê ni Libôkê.*" has the `eng` gloss "*He was living in Libôkê. OR: He used to live in Libôkê.*" To reduce ambiguity at train time, we split each of these double glosses into two separate `eng` sentences. For $\text{Train}_{\text{KMT}}$, we duplicate each Creole sentence and use each `eng` gloss to create two pairs of parallel sentences; for $\text{Validation}_{\text{KMT}}$, we retain only the first gloss as the `eng` translation of each Creole sentence.

Table 1 shows the combined dataset sizes after pre-processing. For complete details on the train, validation, and test splits for each language, including both our data and the organizer-provided data before and after cleaning, see Tables 7, 8 and 9.

	Train	Val.	Test_{KMT}	$\text{Test}_{\text{Ours}}$	All
pap	105,805	1,085	1,967	1,000	109,857
pov	43,699	1,027	33	1,000	45,759
kea	9,438	1,084	163	1,000	11,685
cri	1,376	189	33	155	1,753
pre	105	36	36	0	177
aoa	71	35	39	0	145
fab	61	31	38	0	130

Table 1: Numbers of Parallel Sentences (with `eng`) for each language pair, ordered by size of dataset.

[16] Specifically, those collected from the APiCS data source.

4 Models

To create our MT systems, we fine-tune the three multilingual pre-trained translation models described in Section 2: **KMT** (Robinson et al., 2024), **mBART-50** (Tang et al., 2021), and **NLLB** (NLLB Team et al., 2022). We explain our approach for fine-tuning each model below, listing additional training configuration details in Appendix E.

4.1 Baselines

The baseline models specified by the organisers for the unconstrained track of the Systems Subtask were `CreoleM2M` (Lent et al., 2024) and `kreyol-mt` (Robinson et al., 2024).[17] Both were created by fine-tuning `m2m` mBART-50 (Tang et al., 2021) on private datasets. While `CreoleM2M` performs slightly better than `kreyol-mt` on `pap-eng` and `eng-pap` translation, it does not support our other six Creoles of focus, and so for simplicity we use `kreyol-mt` as our experimental baseline.

4.2 Our Models

Fine-tuned KMT We first explore whether we can improve the performance of the baseline `kreyol-mt` model[18] by fine-tuning it further on our datasets using PyTorch Lightning (Falcon and team, 2019). We use `kreyol-mt`'s existing language tags and embeddings for each Creole.[19] Like mBART-50, `kreyol-mt` has 611M parameters and a SentencePiece (Kudo and Richardson, 2018) vocabulary of 250k subwords.

Fine-tuned mBART-50 We then explore whether we can recreate our own version of `kreyol-mt` by fine-tuning the `m2m` version of mBART-50 on our novel datasets using Fairseq (Ott et al., 2019). As the English-centric `many-to-one (m2o)` and `one-to-many (o2m)` versions of mBART-50 have been shown to outperform their `m2m` counterpart (Liu et al., 2020), we also use these models for fine-tuning. All three mBART-50

[17] While these baselines were listed on the Shared Task website, organisers clarified afterwards that `kreyol-mt` has been trained on portions of text from Test_{KMT}, and that the intended baseline was, in fact, `kreyol-mt-pubtrain`.

[18] We chose to fine-tune `kreyol-mt` without realising that its training data included text from the public Test_{KMT} set. The results of these models on Test_{KMT} are therefore inflated.

[19] We note that `kreyol-mt` was trained with `src` language tags appended to the end of each training `src` sentence (in contrast to traditional mBART-50 language tagging in which the `src` tag is prepended to the beginning of the `src` sentence). We replicate the `kreyol-mt` tagging system for tokenising the training, validation and test data.

models share the same SentencePiece (Kudo and Richardson, 2018) vocabulary of 250k subwords. We repurpose existing language tags for our unseen language pairs following Robinson et al. (2024), initialising their embeddings randomly. To compensate for the imbalance in dataset sizes across languages, we use temperature-based sampling with $\tau = 2$, which increases the relative sampling probability of low-resource languages and promotes more balanced training.

Fine-tuned NLLB As a state-of-the-art translation model designed specifically to perform well on low-resource languages, NLLB (NLLB Team et al., 2022) is also commonly fine-tuned for unseen language pairs in specific translation contexts (Ebrahimi et al., 2023; De Gibert et al., 2025). The largest NLLB model is a 54.5B parameter sparsely-gated mixture of experts model; we use two smaller distilled versions of this model (`distilled-1.3B` and `distilled-600M`) for our experiments. While `pap` and `kea` are already supported in NLLB, we add additional language tags for the other five languages and initialise their embeddings randomly. We use PyTorch Lightning for training as described for fine-tuning `kreyol-mt`, except for fine-tuning NLLB where we implement a maximum of 30 training epochs to keep total training time feasible.

Fine-tuning Experiments We first fine-tune `kreyol-mt`, the three different versions of mBART-50 and the two different versions of NLLB on our dataset for three translation directions; all Creoles into `eng` (`XX-eng`), `eng` into all Creoles (`eng-XX`), and both of these directions simultaneously (`XX-XX`). We select the best overall setup for each of the three base models for translation both into and out of `eng`, and then repeat each of those best setups for the following experiments:

1. Initialising embeddings for Creole language tags with existing embeddings in each model for `por`, instead of using existing Creole embeddings (for `kreyol-mt` models) or random initialisation (for NLLB and mBART-50).[20]
2. Including `eng-por` or `por-eng` as an additional training direction, leveraging the high-quality parallel data collected from Tatoeba (see Section 3.1.2).
3. Using `kreyol-mt` distilled data (see Section 3.2) as target side translations for fine-tuning on `pap`, `kea`, `pov` and `cri`.

For each of these fine-tuned models, we find the checkpoint with the highest scores across all languages on the validation set, and then use this best checkpoint to evaluate that model's performance on Test_{KMT}. Where any two experimental settings show improvements on the basic setup for a given base model, we also combine them together.

4.3 Model merging

To obtain most of our final models we applied model merging using Arcee's MergeKit framework (Goddard et al., 2024), specifically the `linear` method (Wortsman et al., 2022). We define three different merging strategies: (i) averaging different checkpoints of the same training run, (ii) merging different (our) models or (iii) merging our models with the `kreyol-mt` baseline model (i.e. federated learning, as the training set of `kreyol-mt` is not public). While the two first options were applied to fine-tuned mBART-50 and NLLB models (described in Section 4.2), the last option was applied to the fine-tuned KMT models (Section 4.2). In our experiments we merge between 3 and 5 checkpoints, mostly from our internal finetuned models (selecting based on best-performance on the validation dataset for specific language pairs), but also – in the case of (iii) – external models. We note that most of the time, this procedure meant averaging three last checkpoints of our finetuned models.

4.4 Post-editing

With the lexicons we collected for each Creole and the system outputs of the best models for each language pair on the Test_{KMT} dataset, we implement post-editing with three LLMs; Gemini 1.5 Pro, Mistral Large 2.1 and Open AI's GPT 3.5 Turbo.[21] Following Nielsen et al. (2025), our first prompting strategy (P1) includes only the source sentence and the system translation, while our second prompting strategy (P2) includes the translations as well as the entire lexicon for the relevant language pair. For each of these two strategies, we experiment with using the exact prompt proposed in Nielsen et al. (2025) as well as our own prompt construction. All four prompts are listed in full in Table 11 in Section B. For `pap`, we repeat the experiment with both the traditional bilingual lexicon and the GATITOS lexicon (Jones et al., 2023).

[20] For NLLB, as `pap` and `kea` are already supported languages in the pre-trained model, we do not reset the embedding weights for these language tags in the same fashion.

[21] Due to resource limitations, we did not use the paid OpenAI model to post-edit the `pap` Test_{KMT} dataset, which is over ten times as long as the test sets for the other six languages.

ID	XX→eng								eng→XX							
	pap	pov	kea	cri	pre	fab	aoa	all	pap	pov	kea	cri	pre	fab	aoa	all
kreyol-mt	75.1	89.0	94.0	83.1	10.6	11.3	11.0	53.4	66.4	91.8	91.8	80.0	8.38	6.65	8.56	50.5
KMT1	75.4	69.2	**91.1**	73.5	31.8	14.7	**19.9**	**53.7**	68.0	56.4	71.2	64.2	19.4	12.9	17.5	44.2
A. + por embeddings	**75.9**	68.0	89.6	66.3	32.7	15.2	19.0	52.4	66.9	61.4	74.5	45.7	18.7	**14.0**	**17.6**	42.7
B. + por data	75.6	68.7	89.8	67.3	29.7	14.7	19.3	52.2	**67.6**	56.2	70.4	55.2	**21.5**	12.1	**17.6**	42.9
C. + distilled data	73.1	**75.5**	89.7	**80.1**	25.3	13.7	18.0	53.6	65.6	66.6	**84.5**	**75.0**	0.0	0.0	0.0	41.7
D. + A + C	71.8	72.5	86.5	64.5	**36.9**	**15.4**	18.3	52.3	63.0	**71.9**	81.6	52.5	16.6	12.3	15.5	**44.8**
MB1/MB2	76.1	49.6	63.3	33.9	**50.7**	20.8	26.7	**46.2**	**73.1**	32.4	**44.1**	**26.5**	26.4	17.0	**28.3**	**35.4**
A. + por embeddings	**76.4**	50.0	**63.5**	32.9	50.2	20.1	27.3	45.8	**73.1**	33.4	43.1	25.6	27.2	**17.5**	26.0	35.1
B. + por data	75.6	50.4	63.3	34.9	47.7	**22.1**	**27.9**	46.0	71.3	29.6	40.1	21.7	**28.6**	**17.5**	25.2	33.4
C. + distilled data	74.4	50.8	62.2	**36.6**	43.9	20.4	25.2	44.8	71.3	**36.7**	39.1	22.0	23.9	15.3	20.1	32.6
D. + A + B	75.6	**54.0**	63.2	35.9	48.4	19.4	27.1	**46.2**	71.5	29.3	39.5	22.2	24.1	17.1	24.8	32.6
NLLB1/NLLB2	**83.3**	**55.5**	70.5	24.8	35.6	20.4	21.0	**44.4**	77.1	52.5	56.3	28.1	23.4	**18.4**	**24.6**	**40.1**
A. + por embeddings	82.6	51.3	68.2	**27.0**	**39.9**	20.3	20.2	44.2	74.2	49.5	52.5	24.8	24.2	14.9	20.9	37.3
B. + por data	83.1	49.9	68.6	24.0	37.9	**20.7**	**25.1**	44.0	75.5	53.0	56.6	**31.9**	**28.3**	16.2	18.4	40.0
D. + A + B	83.0	49.7	**72.0**	24.6	31.4	20.6	19.5	43.0	**77.3**	**53.8**	**56.7**	26.1	26.0	**18.4**	22.0	40.0

Table 2: Results of fine-tuning experiments (A) initialising language embeddings with por embeddings; (B) adding high-quality por data to training data; (C) using distilled data as training data for pap, kea, pov and cri, and (D) any relevant combinations of the three conditions. Results calculated on Test_{KMT} dataset, using single best checkpoint for each model (as evaluated on validation set). Results in **bold** indicate best results for that language pair out of all experimental settings for that base model; highlighted results are best out of all fine-tuned models (green = beats kreyol-mt baseline).

5 Results and Discussion

In this section, we report and discuss the results of our fine-tuning experiments, model merging and post-editing with LLMs. All results are calculated using the ChrF metric[22] (Popović, 2015) implemented in the SacreBLEU library (Post, 2018).[23]

Fine-tuning We find through initial fine-tuning on our dataset that the best overall models for translation into eng are kreyol-mt fine-tuned for XX-XX translation (KMT1), mBART-50 m2m fine-tuned for XX-eng translation (MB1) and NLLB distilled-1.3B fine-tuned for XX-eng translation (NLLB1). We find the best overall models for translation out of eng are kreyol-mt fine-tuned for XX-XX translation (KMT1), mBART-50 o2m fine-tuned for eng-XX translation (MB2) and NLLB distilled-1.3B fine-tuned for eng-XX translation (NLLB2). For each of these best setups, we then implement our initial set of experiments by retraining each model using por embeddings, por data or distilled data, and then the combinations of the two best settings for each base model.

The results in Table 2 show that different strategies work best for different base models, directions and language pairs – there is no single experimental setting that shows across-the-board advantages. NLLB-based models (NLLB1/NLLB2) show the strongest performance on translation to and from pap, which is not surprising given that this is one of NLLB's supported languages and that the model has seen large amounts of pap data during pre-training. However, using distilled data does not improve the NLLB1/NLLB2 results for pap nor any other language pairs, therefore we exclude the results for this setting. The mBART-50-based models (MB1/MB2) outperform the other fine-tuned models on aoa, fab and pre, except for eng-fab translation. Their high performance on these languages (the lowest-resourced in the set) is likely due to the temperature sampling strategy utilised in our fine-tuning setup for mBART. Conversely, the fine-tuned kreyol-mt model (KMT1) performs better than the other fine-tuned models on kea, pov and cri in both translation directions, particularly when training on distilled data.

Our best model for kea, pov and cri (fine-tuned kreyol-mt) does not beat the kreyol-mt baseline in these languages, so we experiment further with fine-tuning kreyol-mt. We therefore repeat the three experiments while fine-tuning kreyol-mt only for one translation direction at a time (XX-eng or eng-XX), as well as fine-tuning on only the highest-resource languages (cri, pov, kea and pap). To further improve scores, we find each model's best checkpoint for each language pair on the validation set and then use this checkpoint to translate Test_{KMT} for that language pair. Any of these new models which improve on our previous best results for a given language pair are included in Table 13 in Section D, along with the per-language checkpointed scores for the other best models per language pair from Table 2.

[22] Note that we use ChrF but the official Shared Task proceedings uses ChrF++.

[23] nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1

	XX→eng									eng→XX							
ID	pap	pov	kea	cri	pre	fab	aoa	all	ID	pap	pov	kea	cri	pre	fab	aoa	all
kreyol-mt	75.1	89.0	94.0	83.1	10.6	11.3	11.0	53.4		66.4	91.8	91.8	80.0	8.38	6.65	8.56	50.5
H1	73.9	57.4	67.4	36.4	**55.3**	**28.2**	**35.1**	50.5	H4	66.2	67.5	**90.4**	**78.3**	0.0	0.0	0.0	**43.2**
H2	**84.4**	68.3	72.3	37.6	39.5	22.3	21.9	49.6	H5	**77.6**	52.5	57.1	27.3	**27.8**	**18.7**	**25.9**	41.0
H3	76.8	80.9	**93.6**	**82.3**	24.0	12.7	19.1	**55.6**	H6	59.3	**73.7**	76.0	47.0	16.9	10.2	14.3	42.5
H4	73.9	**81.4**	92.9	76.3	22.9	13.7	17.8	54.1									

Table 3: Results of model merging, calculated on Test_{KMT} dataset. Results in **bold** indicate best results for that language pair across all merged models; highlighted results are better than all other fine-tuned models (green = beats kreyol-mt baseline).

For translation into eng, fine-tuning kreyol-mt for XX-eng translation only gave best results for kea-eng and cri-eng (KMT2). Fine-tuning kreyol-mt for XX-XX translation but with distilled data and only with the higher-resource Creoles (pap, pov, kea and kea) improved results for pov-eng translation (KMT3). For translation out of eng, fine-tuning kreyol-mt for eng-XX translation only gave best results for eng-kea and eng-pov, using distilled data for the former (KMT4) and distilled data plus initialisation with por embeddings for the latter (KMT5). Despite these improvements, no models beat kreyol-mt scores for pov, kea and cri in either translation direction; and KMT1C remains our best-performing model for eng-cri.

Model Merging We create a total of six new models by merging different combinations of our fine-tuned and base models. Results across all language pairs and translation directions are displayed in Table 3. To improve performance on the lowest-resource languages (aoa, fab and pre) we first combine the best checkpoints of MB1B (2 checkpoints) and MB1C (3 checkpoints), obtaining the H1 model. For pap-eng we try averaging the last three checkpoints of NLLB1 (H2) and for eng-pap we take the same approach for NLLB2D (H5). For pov, kea and cri, for XX-eng we try averaging the last three checkpoints of KMT2, but find no improvements on our best scores and so exclude this model from our results. For eng-XX we average the last three checkpoints of KMT5 (H6), obtaining a new best-score for eng-pov translation. Finally, we explore whether incorporating the base kreyol-mt model directly in the merging can improve scores, combining the last three checkpoints of KMT2 with kreyol-mt (H3) and the last three checkpoints of KMT1C with kreyol-mt (H4). Our six model merges beat our existing best scores on all language directions except eng-aoa, eng-fab and eng-pre; yet our new best scores for translation from and into kea, pov and cri still do not beat the kreyol-mt baseline.

Post-editing Finally, we take our best models for each language direction and post-edit their Test_{KMT} outputs with different LLMs. We include a full list of results in Table 14 in Section D. In most cases, the LLM-edited outputs are worse than the original system outputs, but we obtain modest improvements for fab-eng, eng-fab, pre-eng, eng-pre and eng-aoa translation. For every translation direction, post-editing with the lexicon gives better results than post-editing without the lexicon, even for aoa which has only a small, hand-crafted lexicon. For pap, we obtain better results using the traditional bilingual lexicon than the GATITOS lexicon, despite the fact that the GATITOS lexicon is over three times larger than the former, potentially indicating that the lexical items included in the former are more useful for this test set domain.

Final models Out of all our finetuning, merging and post-editing experiments, we select the best systems to submit to the Shared Task, reporting the performance of each system on the test set in Table 4. The first submissions are generated by the single best model for XX-eng translation (merged model H3) and eng-XX translation (best overall checkpoint of MB2, a finetuned mBART-50 model).[24] The second submissions are generated by the best models or checkpoints for each individual language pair, except for eng-kea and eng-cri where there is no better model or checkpoint than Submission 1. We also include a third submission for translation directions where the LLM post-editing resulted in improvements on the second submission outputs.

6 Data Contamination

At the end of the Shared Task, Organisers communicated with us that the kreyol-mt model, one of

[24]We selected MB2 because, when evaluated on each language with the best checkpoint per language, it showed the highest average performance across all language directions. However, we realised in hindsight that the best *single* checkpoint across all language pairs was actually from KMT1D.

	XX→eng							eng→XX						
	pap	pov	kea	cri	pre	fab	aoa	pap	pov	kea	cri	pre	fab	aoa
kreyol-mt	75.1	**89.0**	**94.0**	**83.1**	10.6	11.3	11.0	66.4	**91.8**	**91.8**	**80.0**	8.38	6.65	8.56
Sub. 1 (H3/MB2)	76.8	80.9	93.6	82.3	24.0	12.7	19.1	73.1	32.4	44.1	26.5	26.4	17.0	28.3
Sub. 2 (best per LP)	**84.4**	81.4			55.3	28.2	**35.1**	**77.6**	73.7	90.4	78.0	41.7	25.7	33.6
Sub. 3 (Sub. 2 + LLM)					**57.1**	**28.7**						**44.2**	**26.6**	**33.6**

Table 4: ChrF scores for system submissions from best single models per translation direction (Sub. 1), best models per language pair (Sub. 2) and best models per language pair + LLM post-editing (Sub. 3) on the Test_{KMT} dataset (Bold = best score, green highlight = beats kreyol-mt baseline). Unfortunately, XX-eng model outputs for Submission 2 (grey) were not submitted to the Shared Task due to administrative error.

		XX→eng							eng→XX						
		pap	pov	kea	cri	pre	fab	aoa	pap	pov	kea	cri	pre	fab	aoa
$\text{Test}_{\text{KMT-D}}$	kreyol-mt	68.4	42.8	57.9	37.3	6.00	11.0	10.4	60.3	29.7	51.6	27.4	8.93	5.47	9.55
	Submission 1	**67.3**	**50.7**	**61.9**	**39.4**	26.4	21.9	26.7	48.4	27.3	**45.8**	36.0	26.0	**41.2**	**46.2**
	Submission 2	64.4	39.7	-	-	**60.0**	**48.4**	**50.1**	**59.6**	**51.3**	27.4	**40.5**	**26.5**	39.0	31.3
$\text{Test}_{\text{Ours}}$	kreyol-mt	39.5	29.8	-	-	-	-	-	38.8	20.1	-	-	-	-	-
	Submission 1	45.8	28.6	-	-	-	-	-	26.9	**44.2**	-	-	-	-	-
	Submission 2	**67.6**	**46.2**	-	-	-	-	-	**49.5**	18.4	-	-	-	-	-

Table 5: Results for kreyol-mt baseline model compared to our Submission 1 and Submission 2 models on $\text{Test}_{\text{KMT-D}}$ and $\text{Test}_{\text{Ours}}$. Bold = best score, green highlight = beats kreyol-mt baseline.

the specified baseline models for the unconstrained systems track, had been trained on some of the Shared Task public test data; and the intended baseline was kreyol-mt-pubtrain. This explains why kreyol-mt scored so highly on the official test set for certain language pairs (kea, pov and cri), and why our models cannot beat it in these directions despite additional data and modelling efforts.

For our submission, this clarification impacted our experimental baseline and our finetuned or merged models which use kreyol-mt as a base model. This means a substantial proportion of our submissions were affected.[25] To address this, we re-evaluated both the kreyol-mt baseline and our Submission 1 and Submission 2 models[26] on two further test sets:

- A decontaminated version of the KMT test datasets ($\text{Test}_{\text{KMT-D}}$) provided by the organisers, with data not seen during training of either kreyol-mt or kreyol-mt-pubtrain (see dataset sizes in Table 10).
- $\text{Test}_{\text{Ours}}$, which is made of pap and pov data we collected but did not use for training, including data from domains not seen during training of kreyol-mt (see dataset sizes in Table 7).[27]

The results (Table 5) show that our Submission 1 and Submission 2 models outperform kreyol-mt in 12 out of 14 translation directions (all except pap-eng and eng-pap) on $\text{Test}_{\text{KMT-D}}$. On $\text{Test}_{\text{Ours}}$, our Submissions beat kreyol-mt in all four translation directions, including pap-eng and eng-pap. These results provide a more realistic picture of the performance of the baseline and our own models on the different language pairs, without inflation on a contaminated test set. Furthermore, kreyol-mt performs considerably worse on the FLORES benchmark (Goyal et al., 2022) for pap and kea (see Appendix C) than on either Test_{KMT} or $\text{Test}_{\text{KMT-D}}$. These results indicate that, aside from the issue of data contamination, the kreyol-mt model seems to be heavily overfitted to KMT-style data and less good at generalising to novel domains. We note that this may have also degraded the quality of our backtranslated training data, since we use three kreyol-mt models to backtranslate monolingual Creole data from different domains into English (see Section 3.2).

[25] Specifically, our finetuned and merged models which used kreyol-mt as a base model included H3, H4, H5 and H6, used for Submission 1 and Submission 2 for several language pairs.

[26] Due to resource and time constraints, we were not able to repeat our LLM-post editing techniques (creating Submission 3) on the new test sets.

[27] We split out this test data *after* synthetically creating parallel data by using kreyol-mt-pubtrain and kreyol-mt-scratch models to backtranslate monolingual data (see Section 3.2). As a result, 13% and 15% of our pap and pov test sets are made up of synthetic data. We also have our own test data for kea and cri (see Table 7) but because a much higher proportion of these splits are synthetic (63% and 100% respectively), we do not evaluate on this data here.

7 Conclusion

Our submissions to the WMT 2025 Creoles MT Systems Subtask utilise a range of known MT techniques, including fine-tuning three pre-trained multilingual translation models on both task data and additional data, merging best models and checkpoints and post-editing system outputs using LLMs. While no single fine-tuning, merging or post-editing strategy emerged as best amongst all language pairs, we observed considerable gains over the baseline KMT model performance on the Test_{KMT} dataset for `pap`, `aoa`, `fab` and `pre` by combining different approaches, including oversampling the lowest-resource languages in the training data via temperature sampling. While some of our results are unreliable due to the fact that Test_{KMT} is contaminated with `kreyol-mt` training data, we demonstrate the robustness of our model's performance using alternative test sets, and show that `kreyol-mt` appears to be overfitted to KMT-style data in general. Future work could explore whether the techniques and strategies we have utilised here to improve performance are also useful for other Creole language pairs and across data from a broader variety of different domains.

Limitations

The official Shared Task test sets for these languages are identical to the test sets which are publicly available on Hugging Face, meaning that the gold labels were available at the point of submission. We ensured that no samples from these test sets were in our own training data. However, before we realised that the official test set would be identical to the public one, we made modelling and design decisions based on performance on the publicly-available test set. For example, we selected the best of the four `kreyol-mt`, `kreyol-mt-pubtrain`, `kreyol-mt-scratch` and `kreyol-mt-pubtrain-scratch` models for forward translation and backward translation of our training data based on their performance on the publicly available test set, both per language and overall. We also selected our models for submission based on their performance on this test set, given that the gold labels were freely available. This biases our model development process towards this particular test set, potentially reducing generalisability or robustness of the overall MT systems and potentially giving us an advantage in the context of the Shared Task.

A key limitation of our work is that our modelling decisions and comparisons were initially guided by the `kreyol-mt` model, which was mistakenly announced as the Shared Task baseline. The organisers later clarified that this model had been trained on portions of the Test_{KMT} set, meaning not only that the baseline we were comparing to was trained on the data we were testing on, but also that our models which use it as a base model are also likely inflated. We address this in Section 6 but reiterate here that the results for our KMT-based models in Table 2, and the results for H3, H4, H5 and H6 in Table 3 and Table 4 are likely inflated.

In addition, `kreyol-mt` was trained using a non-standard tagging scheme, appending `src` language tags to the end of source sentences rather than prepending them as in standard mBART-50. Our models inherit this convention, which may limit comparability with other mBART-based systems.

Acknowledgments

This research was supported by the UK Research and Innovation (UKRI) AI Centre for Doctoral Training in Designing Responsible Natural Language Processing (grant number EP/Y030656/1); the National Science Centre, Poland (2023/49/N/ST6/02691); the EU Horizon Europe Research and Innovation programme (GA No 101070350) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (GA No 10052546); the OpenEuroLLM project, co-funded by the Digital Europe Programme (GA No 101195233); and EU Horizon Europe Research and Innovation programme (GA No 101070631) and UK Research and Innovation under the UK government's Horizon Europe funding guarantee (GA No 10039436).

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2025/018209. We also thank Bhavitvya Malik of the StatMT group at the University of Edinburgh for sharing his code for fine-tuning NLLB with PyTorch Lightning, which we adapted for our experiments.

References

Vanessa Pinheiro de Araújo and Gabriel Antunes de Araujo. 2013. Um dicionário principense-português. Master's thesis, Faculdade de Filosofia, Letras e Ciências Humanas, University of São Paulo.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Raj Dabre and Aneerav Sukhoo. 2022. KreolMorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Raj Dabre, Aneerav Sukhoo, and Pushpak Bhattacharyya. 2014. Anou tradir: Experiences in building statistical machine translation systems for mauritian languages – creole, English, French. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88, Goa, India. NLP Association of India.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

William Falcon and The PyTorch Lightning team. 2019. PyTorch Lightning. `https://zenodo.org/records/3828935`. Accessed: 2025-08-06.

Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptiste, Raj Dabre, Johannes Bjerva, and Heather Lent. 2025. Limited-resource adapters are regularizers, not linguists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 222–237, Vienna, Austria. Association for Computational Linguistics.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tjerk Hagemeijer, Philippe Maurer-Cecchini, and Armando Zamora Segorbe. 2020. *A Grammar of Fa d'Ambô*. De Gruyter Mouton, Berlin, Boston.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.

Jonathan Hus, Antonios Anastasopoulos, and Nathaniel Krasner. 2025. Machine translation using grammar materials for LLM post-correction. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 92–99, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics*, 9:790–806.

Heather Lent, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard. 2021. On language models for creoles. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 58–71, Online. Association for Computational Linguistics.

Heather Lent, Emanuele Bugliarello, and Anders Søgaard. 2022. Ancestor-to-creole transfer is not a walk in the park. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Rao Ma, Mengjie Qian, Yassir Fathullah, Siyuan Tang, Mark Gales, and Kate Knill. 2025. Cross-lingual transfer learning for speech translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 33–43, Albuquerque, New Mexico. Association for Computational Linguistics.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Elizabeth Nielsen, Isaac Rayburn Caswell, Jiaming Luo, and Colin Cherry. 2025. Alligators all around: Mitigating lexical confusion in low-resource machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 206–221, Albuquerque, New Mexico. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International*

Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Nathaniel Robinson, Cameron Hogan, Nancy Fulda, and David R. Mortensen. 2022. Data-adaptive transfer learning for translation: A case study in Haitian and jamaican. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 35–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Raj Dabre, Kenton Murray, Andre Coy, and Heather Lent. 2025. Findings of the first shared task for creole language machine translation at wmt25. In *Proceedings of the Tenth Conference on Machine Translation*.

Jacqueline Rowe, Edward Gow-Smith, and Mark Hepple. 2025. Limitations of religious data and the importance of the target domain: Towards machine translation for Guinea-Bissau creole. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 183–200, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *Preprint*, arXiv:2201.03110.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23965–23998. PMLR.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

A Data Collection

Data Type	L1	L2	Description	Source	No. items
Parallel	por	eng	Tatoeba Translation Challenge	Tiedemann (2020)	112,376
	pap	eng	Bible data	Bible.com	29,367
	pap	eng	Watchtower Series†	The Jehovah's Witnesses	4,275
	pap	eng	Online Random Sentence Generator	Sapaté, na bo sapatu!	5,936
	pov	eng	Bible data	Bible.com	29,876
	pov	eng	Watchtower Series	The Jehovah's Witnesses	8,685
	pov	por	Bilingual dictionary gloss sentences	Dicionário Bilíngue	1,603
	pov	eng	Article on internet access	Open Global Rights	18
	kea	eng	Watchtower Series	The Jehovah's Witnesses	4,273
	fab	eng	Translated stories	Hagemeijer et al. (2020)	430
	pre	por	Bilingual dictionary gloss sentences	Araújo and Araujo (2013)	81
	aoa	eng	IMT Vault sentences	IMT Vault	46
Monolingual	pov	-	Monolingual dictionary gloss sentences	Amarilio Da Mata*	6,930
	pov	-	Documentary Subtitles	Language and Society in Guinea-Bissau	254
	pov	-	Song Lyrics	Tino Trimó via Letras	177
	pap	-	Song Lyrics	Lyrics Translate‡	5,803
	kea	-	School Textbook	Língua e Cultura Cabo-verdiana 10º ano	2,688
	kea	-	Blogposts	Odju d'Agu	2,357
	kea	-	Song Lyrics	Cesária Évora via Letras	2,317
	cri	-	Watchtower Magazine	The Jehovah's Witnesses	1,554
Lexical	cri	por	Bilingual Lexicon	Dicionário livre santome/português	4,929
	pap	eng	Bilingual Lexicon	GATITOS	4,001
	pap	eng	Bilingual Lexicon	Parleremo	1,307
	pov	por	Bilingual Lexicon	Dicionário Bilíngue	1,983
	kea	eng	Bilingual Lexicon	Disonariu Kabuverdianu	1,763
	pre	por	Bilingual Lexicon	Araújo and Araujo (2013)	1,684
	fab	eng	Bilingual Lexicon	Hagemeijer et al. (2020)	473
	aoa	eng	Bilingual Lexicon	IMT Vault§	68

†Corsou dialect.
‡ We collected only lyrics which were tagged exclusively with the pap language tag and no other language tags.
* This is an unpublished manuscript shared privately with the lead author. Lexical items and their definitions were made into full sentences for the purposes of model training by appending each lexical item + 'i' *(is)* + definition.
§ For aoa, we could not find an official lexicon and therefore manually curated a small set of parallel lexical items using the word-aligned entries in the IMT Vault resource.

Table 6: Raw data sources and sizes. Rows shaded in gray were collected too late in the Shared Task period for us to use for model training, but are included here in case useful for future research.

	Train	$\text{Train}_{\text{Clean}}$	Validation	Test	All†	avg. length
pov	44,275	43,419	1,000	1,000	45,419	26.2
pap	43,381	40,850	1,000	1,000	42,850	23.3
kea	8,501	8,099	1,000	1,000	10,099	27.5
cri	1,244	1,218	155	155	1,528	14.4
aoa	0	0	0	0	0	0
fab	0	0	0	0	0	0
pre	0	0	0	0	0	0

†Calculated using cleaned training data.

Table 7: Numbers of parallel sentences for each language pair from **our** data, ordered from highest to lowest resourced. For training data, we show the numbers of raw and cleaned sentences (e.g. after pre-processing). Average length is calculated as average number of words per sentence across all data splits.

	Train	$\text{Train}_{\text{Clean}}$	Validation	Test	All†	avg. length
pap	65,094	64,983	85	1,967	67,035	22.1
kea	1,470	1,340	84	163	1,587	16.6
pov	389	284	27	33	344	5.8
cri	209	155	34	33	222	6.0
pre	147	105	36	36	177	5.8
fab	109	61	31	38	130	5.4
aoa	99	71	35	39	145	6.5

†Calculated using cleaned training data.

Table 8: Numbers of parallel sentences for each language pair from **Organiser-Provided** data, ordered from highest to lowest resourced. For training data, we show the numbers of raw and cleaned sentences (e.g. after pre-processing). Average length is calculated as average number of words per sentence across all data splits.

	Train	$\text{Train}_{\text{Clean}}$	Validation	Test	All†	avg. length
pap	108,475	105,698	1,085	2,967	109,750	22.6
pov	44,664	43,701	1,027	1,033	45,761	26.1
kea	9,971	9,439	1,084	1,163	11,686	26.0
cri	1,453	1,375	189	188	1,752	13.4
pre	147	105	36	36	177	5.8
fab	109	61	31	38	130	5.4
aoa	99	71	35	39	145	6.5

†Calculated using cleaned training data.

Table 9: Numbers of parallel sentences for each language pair from **our and Organiser-Provided** data, ordered from highest to lowest resourced. For training data, we show the numbers of raw and cleaned sentences (e.g. after pre-processing). Average length is calculated as average number of words per sentence across all data splits.

	Test	avg. length
pap	1,896	17.9
pov	23	2.9
kea	34	15.2
cri	33	4.8
pre	36	3.7
fab	34	6.8
aoa	35	6.2

Table 10: Numbers of parallel sentences for each language pair from the **Decontaminated Organiser-Provided Test set**, ordered from highest to lowest resourced. Average length is calculated as average number of words per sentence across all data splits.

B Prompts used for LLM post-editing

Condition	Nielsen et al. 2025	Ours
1. Post-editing without lexicon	**P1A:** You are asked to edit the following translation from {src_code} into {tgt_code}. The proposed translation is high-quality, but may have some incorrect words. Please output only the translation of the text without any other explanation. {src_code}: {source} {tgt_code}: {model_translation}	**P1B:** You are given a source sentence and a translation. Improve the translation from {src_code} into {tgt_code}. You must return ONLY the corrected translation sentence, without explanation or extra text. Source: {source} Translation: {model_translation}
2. Post-editing with lexicon	**P2A:** You are asked to edit the following translation from {src_code} into {tgt_code}. The proposed translation is high-quality, but may have some incorrect words. Note the following translations: Lexicon: {lexicon_str} Please output only the translation of the text without any other explanation. {src_code}: {source} {tgt_code}: {model_translation}	**P2B:** You are given a source sentence, a translation and a lexicon. Improve the translation from {src_code} into {tgt_code}. You must return ONLY the corrected translation sentence, without explanation or extra text. Source: {source} Translation: {model_translation} Lexicon: {lexicon_str}

Table 11: Prompts used in LLM post-editing experiments.

C FLORES Evaluation

Model	pap→eng		eng→pap		kea→eng		eng→kea	
	Test_{KMT}	FLORES	Test_{KMT}	FLORES	Test_{KMT}	FLORES	Test_{KMT}	FLORES
kreyol-mt-pubtrain	79.84	54.39	69.94	60.14	80.66	45.65	52.54	52.16
kreyol-mt	75.10	63.12	66.39	57.27	93.94	55.46	91.76	52.33
kreyol-mt-scratch-pubtrain	74.68	47.17	69.36	55.54	70.23	37.22	49.46	46.98
kreyol-mt -scratch	71.82	60.73	67.19	55.06	89.85	50.83	81.67	49.04
nllb-200-distilled-600M	46.50	59.18	53.18	50.09	59.36	63.04	38.27	41.67
nllb-200-1.3B	58.40	68.88	56.58	55.08	62.68	65.86	41.09	43.02
nllb-200-distilled-1.3B	55.30	69.20	58.02	55.40	59.28	64.89	39.75	42.09
nllb-200-3.3B	60.90	69.16	58.78	55.66	63.69	67.46	43.92	45.76

Table 12: ChrF scores for each `kreyol-mt` model across language directions, evaluated on both Test_{KMT} and FLORES test sets.

D Model Results

		kreyol-mt	kreyol-mt-pubtrain	Ours Best	Model ID	Base model	Fine-tuning Direction	Additional setup
XX-eng	pap	75.1	79.8	83.3	$NLLB1_{pap}$	NLLB 1.3B	XX-eng	-
	kea	94.0	80.7	92.3	$KMT2_{kea}$	KMT	XX-eng	-
	pov	87.8	63.4	78.4	$KMT3_{pov}$	KMT	XX-XX	distilled data + HRLs Only
	aoa	10.9	17.0	34.8	$MB1C_{aoa}$	mBART-50 m2m	XX-eng	distilled data
	cri	83.1	31.7	80.5	$KMT2_{cri}$	KMT	XX-eng	-
	fab	11.3	13.7	27.9	$MB1B_{fab}$	mBART-50 m2m	XX-eng	por data
	pre	10.6	16.2	55.0	$MB1B_{pre}$	mBART-50 m2m	XX-eng	por data
	all	53.2	43.2	55.0	$KMT2_{all}$	KMT	XX-eng	-
eng-XX	pap	66.4	70.0	77.3	$NLLB2D_{pap}$	NLLB15 1.3B	eng-XX	por embeddings + por data
	kea	91.8	52.5	86.2	$KMT4_{kea}$	KMT	eng-XX	distilled data
	pov	91.8	51.6	72.8	$KMT5_{pov}$	KMT	eng-XX	por embeddings + distilled data
	aoa	8.6	13.6	33.6	$MB2B_{aoa}$	mBART02m	eng-XX	por data
	cri	80.0	32.1	78.2	$KMT1C_{cri}$	KMT	XX-XX	distilled data
	fab	6.7	9.3	25.9	$MB2_{fab}$	mBART-50 o2m	eng-XX	-
	pre	8.38	10.7	41.7	$MB2_{pre}$	mBART-50 o2m	eng-XX	-
	all	50.5	34.3	46.1	$MB2_{all}$	mBART-50 o2m	eng-XX	-

Table 13: Settings and results of best-performing model checkpoints for each language. Results are calculated on $Test_{KMT}$ dataset, using the best model checkpoint per language pair based on performance on the validation dataset, as indicated with subscript. For evaluation of all translation directions, we report the models with the best average scores using the best checkpoints for each language pair. New models not previously included in Table 2 are highlighted in gray. Green = beats kreyol-mt and kreyol-mt-pubtrain baselines.

		XX→eng								eng→XX							
	Prompt	pap	kea	pov	aoa	cri	fab	pre	all	pap	kea	pov	aoa	cri	fab	pre	all
Submission 2 models	-	84.4	93.6	80.9	35.1	82.3	28.2	55.3	65.7	77.6	90.4	73.7	33.6	78.0	25.9	41.7	60.1
GPT 3.5 Turbo	P1A	–	76.2	54.0	33.4	50.4	25.2	46.5	47.6	–	74.0	46.2	30.9	56.0	24.9	34.8	44.5
	P1B	–	74.2	55.1	31.5	51.1	25.4	47.4	47.5	–	68.7	41.6	29.6	50.3	25.2	32.4	41.3
	P2A	–	**87.1**	**69.2**	31.2	**78.3**	**28.7**	**51.7**	**57.7**	–	**83.2**	**59.8**	32.5	**75.2**	**25.7**	**41.8**	**53.0**
	P2B	–	81.1	63.5	**32.3**	62.9	24.6	46.6	51.8	–	75.2	55.8	**33.2**	67.0	25.1	40.4	49.5
Mistral Large 2.1	P1A	79.4	84.0	57.2	31.6	57.2	26.2	48.3	54.9	71.5	82.6	47.6	31.6	67.0	23.5	36.0	51.4
	P1B	78.1	81.1	56.2	33.0	55.6	25.2	44.3	53.4	70.6	79.0	46.8	31.5	61.3	25.0	36.7	50.1
	P2A	**83.1**	**91.3**	**79.7**	29.8	**66.1**	24.0	**57.1**	**61.6**	74.1	**88.9**	**61.6**	32.1	64.2	25.8	**42.4**	55.6
	P2B	81.7	85.7	68.6	**34.2**	57.8	**28.1**	51.1	58.2	**76.0**	88.3	58.9	**32.8**	**74.7**	**26.6**	**41.9**	**57.9**
	$P2A_{GAT}$	71.7	–	–	–	–	–	–	–	59.5	–	–	–	–	–	–	–
	$P2B_{GAT}$	70.8	–	–	–	–	–	–	–	66.2	–	–	–	–	–	–	–
Gemini 1.5 Pro	P1A	83.1	84.3	58.4	30.7	52.0	26.6	50.7	55.1	74.0	75.5	46.3	24.3	46.1	22.8	29.2	45.5
	P1B	78.8	75.1	47.8	27.5	44.2	23.7	40.0	48.2	70.0	63.5	38.4	24.5	37.7	25.1	26.8	40.8
	P2A	**83.2**	**86.0**	**66.3**	**32.7**	**57.5**	**27.3**	**53.8**	**58.1**	**75.2**	79.2	48.7	30.5	49.8	**26.0**	**44.2**	50.6
	P2B	81.7	84.4	57.2	31.4	48.1	26.4	45.0	53.4	74.0	**82.7**	**52.7**	**33.6**	**62.2**	25.8	41.3	**53.2**
	$P2A_{GAT}$	82.4	–	–	–	–	–	–	–	73.8	–	–	–	–	–	–	–
	$P2B_{GAT}$	80.1	–	–	–	–	–	–	–	72.0	–	–	–	–	–	–	–

Table 14: Results from post-editing best model outputs with three LLMs. P1 is post-editing without lexicon and P2 is post-editing with lexicon (see Table 11). Baseline scores are from models of Submission 2 for each language pair (Table 4). Results in **bold** are best results for each LLM for each language pair; highlighted results = best out of all LLMs (green = also beats Submission 2 baselines). We do not apply post-editing for GPT 3.5 Turbo for pap (which has an extremely large test set) due to resource constraints.

Submitted models Table 15 documents which models we use to generate our Shared Task submissions:

- For Submission 1, we select the single best model for XX-eng translation (H3) and eng-XX translation (best overall checkpoint of MB2). We selected MB2 because, when evaluated on each language with the best checkpoint per language, it showed the highest average performance across all language directions. However, we realised in hindsight that the best *single* checkpoint across all language pairs was actually from KMT1D.
- For Submission 2, where a model has multiple checkpoints we submit the best checkpoint for that language pair, as indicated with subscripts (except for eng-kea and eng-cri where there is no better model or checkpoint than Submission 1).
- For Submission 3 we submit the best system outputs after post-editing with LLMs when this showed improvements on Submission 2. We indicate which LLM and which prompting strategy (see Table 11) was applied in parentheses.

Due to administrative error, our Submission 2 models for the XX-eng direction were not submitted to the official Shared Task.

		Sub. 1	Sub. 2	Sub. 3
XX-eng	pap	H3	H2	
	pov	H3	H4	
	kea	H3	H3	
	cri	H3	H3	
	pre	H3	H1	+ Mistral (P2A)
	fab	H3	H1	+ GPT (P2A)
	aoa	H3	H1	
eng-XX	pap	MB2	H5	
	pov	MB2	H6	
	kea	MB2	H4	
	cri	MB2	H4	
	pre	MB2	$MB2_{pre}$	+ Gemini (P2A)
	fab	MB2	$MB2_{fab}$	+ Mistral (P2B)
	aoa	MB2	$MB2B_{aoa}$	+ Gemini (P2B)

Table 15: Model IDs for final system submissions.

E Fine-tuning Hyperparameters

KMT & NLLB We fine-tune KMT & NLLB models using PyTorch Lightning (Falcon and team, 2019) on a single GH200 GPU (bf16). We set the batch size to 32, use the Adam optimizer (Kingma and Ba, 2015) with a learning rate $5e-5$, a warm-up phase of 500 updates and maximum training length of 30 epochs. The model performance is validated using ChrF every 5,000 steps, early stopping after three consecutive validations with no improvement in ChrF score.

mBART We fine-tune mBART-50 using fairseq (Ott et al., 2019) with a multi-gpu (4 A100 GPUs, fP1A6). The data loader has used temperature-based sampling ($\tau = 2$). We set the batch size to maximum of 1024 tokens, use the Adam optimizer with a learning rate $3e{-}5$, a warm-up phase of 2500 updates and maximum training length of 40,000 updates. Moreover, we applied label smoothing with $\epsilon_{ls} = 0.2$, dropout of 0.3, and attention dropout of 0.1. The three best checkpoints were retained according to validation performance (based on the validation loss value), with early stopping after 10 validation intervals.

KozKreolMRU WMT 2025 CreoleMT System Description: Koz Kreol: Multi-Stage Training for English–Mauritian Creole MT

Hemkeshsing Y. Rajcoomar
Independent Researcher
yush2398@live.com

Abstract

Mauritian Creole (Kreol Morisyen), spoken by approximately 1.5 million people worldwide, faces significant challenges in digital language technology due to limited computational resources. This paper presents "Koz Kreol," a comprehensive approach to English-Mauritian Creole machine translation using a three-stage training methodology: monolingual pretraining, parallel data training, and LoRA fine-tuning. We achieve state-of-the-art results with 28.82 BLEU score for EN→MFE translation, representing a 74% improvement over ChatGPT-4o. Our work addresses critical data scarcity through use of existing datasets, synthetic data generation, and community-sourced translations. The methodology provides a replicable framework for other low-resource Creole languages while supporting digital inclusion and cultural preservation for the Mauritian community. This paper consists of both a systems and data subtask submission as part of a Creole MT Shared Task.

1 Introduction

Mauritian Creole [1] is spoken by individuals from Mauritius, Rodrigues, Agalega and the Chagos Archipelago. Over the course of its history, Mauritius was visited by the Arabs, colonized by the Dutch, French and the British. Originally, it is a language made of French and Afro-Malagasy languages which was used as a means of communication between slaves and their French masters (Piat, 1999). Over time, as the English rule began and indentured labourers arrived from India, more words infiltrated the existing Mauritian Creole lexicon. With this deep diversity of linguistic families, Mauritian Creole has words that can etymologically be traced back to France, England, Madagascar and both north and south India (Eriksen, 2007). A defining characteristic of creole languages is their dynamic lexicon, which often exhibits clear phonetic and semantic shifts from their source languages (Kouwenberg and Singler, 2009). For example, in Mauritian Creole, kalamindas = candy floss. However, the exact etymology of the word "kalamindas" is unknown. This multilingual substrate influence is characteristic of Creole formation processes, where multiple source languages contribute to the emerging Creole's lexicon and structure (DeGraff, 2001).

Mauritian Creole is considered to be a low-resource language since it lacks digital computational resources for language technology applications (Lent et al., 2022). Despite having a vibrant community of speakers, Mauritian Creole, like many Creole languages, faces social stigmatization and is often perceived as linguistically inferior or underdeveloped compared to its lexifier languages.(Kouwenberg and Singler, 2009).Throughout the years, several efforts have been made by the government of Mauritius to enforce Mauritian Creole as a language rather than a dialect, part of a broader movement for creole language recognition and standardization (DeGraff, 2005). Mauritian Creole has become part of the school curriculum when teaching languages at an early age. However since most Mauritian Creole media are through traditional sources like newspapers or magazines, few digital resources exist. Creoles are generally under represented in language research since they are generally low resource languages whose datasets are seldom publicly available. This digital divide creates significant barriers for the Mauritian community's participation in the modern digital economy and limits access to language technologies that could support cultural preservation and digital inclusion. (Team et al., 2022)

This work addresses these challenges by developing "Koz Kreol," a comprehensive machine translation system for English-Mauritian Creole translation. We present a three-stage training

[1] Also referred to as "Kreol Morisyen"

Proceedings of the Tenth Conference on Machine Translation, pages 1183–1190

methodology combining monolingual pretraining, parallel data training, and Low Rank Adaptation (LoRA) fine-tuning that achieves state-of-the-art performance on this language pair. Our approach strategically combines existing datasets from previous research efforts with high-quality community-sourced translations and synthetic data generation. The resulting system not only advances the state of machine translation for Mauritian Creole but also provides a replicable framework for developing MT systems for other low-resource creole languages, contributing to broader efforts in digital language inclusion and cultural preservation. This paper consists of both a systems and data subtask submission as part of a Creole MT Shared Task (Robinson et al., 2025).

2 Related Work

Low Resource Machine Translation (LRMT) has evolved from early transfer learning approaches (Zoph et al., 2016) and backtranslation techniques (Sennrich et al., 2016) to sophisticated multilingual pre-trained models like mBART (Liu et al., 2020), which achieved up to 12 BLEU[2] (Papineni et al., 2002) points improvement for low-resource pairs[3]. The paradigm for low-resource languages was further established by mBART-50 (Tang et al., 2020), which scaled multilingual pre-training to 50 languages and became the standard approach for many low-resource translation tasks. Early work by Tanzer et al. (2024) established benchmarks for learning translation from minimal linguistic resources, demonstrating how grammar books alone can provide sufficient structural information for basic translation capabilities in truly low-resource scenarios.

Creole languages present distinctive challenges beyond typical low-resource scenarios due to their genealogical complexity, orthographic variability, and historical stigmatization, requiring multi-source transfer learning rather than conventional single-source approaches. Recent creole MT research has made substantial progress across multiple fronts. Dabre and Sukhoo (2022) established foundational baselines with KreolMorisienMT, creating the first comprehensive parallel corpus for Mauritian Creole with 21,810 sentence pairs and demonstrating effective transfer learning from pre-trained multilingual models. Robinson et al. (2024) dramatically scaled Creole coverage with Kreyòl-MT, presenting 14.5 million unique creole sentences across 41 languages supporting 172 translation directions. Lent et al. (2024) introduced CreoleVal, the first comprehensive benchmark spanning 8 NLP tasks across 28 creole languages. Fekete et al. (2025) explored parameter-efficient approaches through adapter architectures for cross-lingual transfer, while Adelani et al. (2022) demonstrated that strategic fine-tuning of large pre-trained models with small amounts of high-quality data can achieve significant improvements.

3 Dataset Construction and Methodology

3.1 Data Sources

3.1.1 Existing Parallel Corpora

High-quality Mauritian Creole data is scarce, particularly parallel translations. Our training data includes monolingual and bilingual resources from KreolMorisienMT (Dabre and Sukhoo, 2022) and parallel bitext from Kreyol-MT (Robinson et al., 2024), mostly drawn from translated Bibles and local dictionaries, totaling around 40K bilingual sentences of generally acceptable quality[4].

The monolingual data was downsampled to 18,145 sentences (~500K tokens), as empirical testing showed this size outperformed larger sets (250K, 1M, 2M tokens), likely mitigating catastrophic forgetting (McCloskey and Cohen, 1989) and avoiding repetitive degeneration (Holtzman et al., 2020) during pretraining.

3.1.2 Synthetic Data Generation

Although our primary goal was to fine-tune an LLM for machine translation, we enriched the training data with greater diversity and nuance by including 2,023 Massively Multilingual Language Understanding (MMLU) questions, 961 Question Answering (QA) items, 692 Topic Classification sentences, and 1,225 grammar prompts. We also enhanced Claude's (Anthropic, 2024) context with 150 high-quality parallel bitexts from Flores Dev and created grammar exercises using Gramer Kreol Morisien (Carpooran, 2005). The resulting synthetic dataset comprises 4,901 sentences.

3.1.3 Community Sourced Bitext

To address the shortage of high-quality parallel data for Mauritian Creole, we launched a community-

[2]BLEU: Bilingual Evaluation Understanding

[3]Translation pairs where atleast one language is low resource

[4]Not grammatically consistent throughout.

driven data collection initiative using a web-based annotation platform. Native speakers contributed English–Mauritian Creole translations in both directions, based on Claude-generated English sentences containing at least 15 words to ensure sufficient context and complexity. A two-stage validation process ensured quality, with each translation reviewed by another native speaker. This effort yielded approximately 300 high-quality parallel sentence pairs to supplement our training corpus.

3.1.4 FLORES-200

FLORES-200 extends the original FLORES-101 benchmark by incorporating 200 languages with comprehensive evaluation datasets, providing standardized dev and devtest splits of approximately 1,000 sentences each for multilingual machine translation evaluation. The FLORES data for Mauritian Creole was sourced through a rigorous translation process involving two qualified native speakers who translated the English sentences into Mauritian Creole. Each translated sentence underwent review by the other translator, ensuring high linguistic accuracy and cultural authenticity through this collaborative validation approach.

The resulting datasets comprise 997 sentences in the dev split and 1,012 sentences in the devtest split, providing a total of 2,009 high-quality parallel sentence pairs for English-Mauritian Creole translation. Given the extremely scarce number of high quality parallel bitext available for Mauritian Creole, our final model underwent finetuning on both the dev and devtest portions to maximize the utilization of these linguistic resources.

3.1.5 Evaluation Dataset

Since we fine-tune on the Flores-200 Devtest, we created a 100-sentence evaluation set to monitor BLEU (Papineni et al., 2002) and ChrF[5] (Popović, 2015) during training and fine-tuning. The hold-out test data was sourced from LALIT[6] newspaper, focusing on global geopolitics to assess performance on news content. Source sentences in English were translated into Mauritian Creole, with only sentences over 15 words included. Each translation was validated by another fluent native speaker. Aware of domain-specific evaluation limitations, we report Flores-200 Devtest results in the appendix (Table 3) where the fine-tuned model uses only Flores-200 Dev.

[5]Character F-score.

[6]lalitmauritius.org

English–Kreol Morisien					
Split	**L**	**AL-en**	**AL-mfe**	**U-en**	**U-mfe**
train	46,160	7.3	6.7	31,195	32,106
dev	997	21.0	21.4	6,695	6,195
devtest	1,012	21.6	21.9	7,054	6,413
lalit (test)	102	48.5	45.2	1,874	1,762
Kreol Morisien Monolingual					
Split	**L**	**AL**	**–**	**U**	**–**
mono	18,145	87.13	–	27,967	–

Table 1: Dataset statistics. L: total sentences/pairs; AL-en/AL-mfe: average word counts for English/Mauritian Creole; U-en/U-mfe: unique word counts for English/Mauritian Creole.

3.2 Dataset Statistics

Table 1 presents comprehensive statistics for our datasets across different splits[7]. Our training dataset exhibits diverse characteristics across different data sources. The training split contains a substantial number of single-word entries representing 1-1 translations sourced from previous lexical datasets sourced by Dabre and Sukhoo (2022), contributing to the lower average word counts (7.3 for English, 6.7 for Mauritian Creole) compared to the evaluation sets.

The FLORES evaluation sets show significantly higher average word counts (21.0-21.6 for source, 21.4-21.9 for target), likely reflecting the more complex sentence structures typical of the FLORES benchmark. With higher average word counts of 48.5 for English and 45.2 for Mauritian Creole, we assume the hold-out test set to have even higher linguistic complexity, consistent with the discourse of news content covering global geopolitics.

4 Experiments

4.1 Experimental Design

In our comprehensive arsenal, we now have an extensive collection of resources including monolingual Creole data, valuable parallel bitext published by previous researchers, the robust Flores-200 training data, and carefully generated synthetic data. Additionally, we have meticulously curated high quality parallel sentences sourced directly from native local speakers through an ambitious community outsourcing project we launched around a year ago. This community-driven approach ensures authentic linguistic representation and cultural ac-

[7]More details in Section B of the Appendix.

curacy in our training data. We will break our carefully designed training recipe down to three distinct important steps: Pretraining, Training and Finetuning. For this comprehensive study, we will use Llama 3.1-8B model and tokenizer as our robust backbone LLM here.

4.2 Training Setup

According to the findings of Xu et al. (2024), there's significant and demonstrable benefit in pretraining a large language model with a language it is previously unfamiliar with. This crucial step helps the model build a rich internal vocabulary, as well as, develop a deep understanding of the intricate semantics of a language. However, this process has to be done extremely carefully and with precise control since it can lead to the detrimental phenomenon of catastrophic forgetting (McCloskey and Cohen, 1989) when fed too much overwhelming data. To maintain this delicate balance, we use a carefully measured 500K monolingual tokens of authentic Mauritian Creole, complemented by 100K tokens of English and French each. We employ full-weight finetuning for this critical foundational portion.

For the next step in our pipeline, we use our extensive parallel bitext sourced by other dedicated researchers as well as our synthetic data, comprising around 46K sentences. These valuable sources are mostly drawn from the carefully translated Bible and comprehensive local dictionaries. Remarkably, one single pass of the data onto a powerful 3.1-8B Llama backbone is already sufficient to see vast and encouraging improvements in translation performance. Training the model with a learning rate of 1e-5 and the AdamW (Loshchilov and Hutter, 2019) optimizer, we stop after 2 complete epochs to prevent overfitting.

For the final and most refined step in our training methodology, we use our mix of Flores 200 dev and devtest sets, for efficient Low-Rank Adaptation (LoRA) (Hu et al., 2021) Finetuning for a maximum of 3 epochs. We conduct a hyperparameter sweep over the critical rank parameter, the scaling factor named "alpha" and the target modules. The LoRA hyperparamters for our best performing results based on BLEU and ChrF metrics are $\alpha = 8$, $r = 16$, target modules = query and value projections.

5 Results

For the sake of our experiment, since Mauritian Creole data is scarce, we are training on "devtest" and we use the LALIT test set for evaluation.

5.1 Baseline Comparisons

Model / Setup	BLEU	CHRF
EN → MFE		
Zero-Shot (Llama 3.1-8B)	4.22	35.37
ChatGPT 4o	16.55	53.58
Mono Only	22.54	51.67
Mono + Train	26.76	59.55
Mono + Train + LoRA	**28.82**	**60.86**
MFE → EN		
Zero-Shot (Llama 3.1-8B)	28.4	57
ChatGPT 4o	**46.63**	**71.22**
Mono Only	43.32	69.23
Mono + Train	41.78	68.60
Mono + Train + LoRA	43.14	70.21

Table 2: BLEU and ChrF scores for different model configurations in EN ↔ MFE translation.

From Table 2, we observe a significant improvement in translation performance when incorporating 500K monolingual tokens into the model. However, the model still lacks the ability to translate the language effectively from English to Mauritian Creole. The training portion of our approach provides the largest performance boost, with a 18.7% increase in BLEU score and an 15.2% increase in ChrF score. Following this stage, we perform fine-tuning using Low-Rank Adaptation (LoRA) using the peft package. (Mangrulkar et al., 2022) Fine-tuning on 2,000 sentences in both translation directions contributes to approximately 10% improvement in BLEU score.

When examining the reverse direction (MFE → EN), the model performs significantly better than in the forward direction, a common finding when translating Low Resource Langauges to English (Neubig and Hu, 2018). The model performs better on BLEU score after passing in 500K tokens of Mauritian Creole only sentences, and the performance slightly declines when training. The improvement from LoRA fine-tuning is more modest compared to training, likely because the model has already achieved strong performance in back-translation and is approaching a performance plateau.

When comparing our model to a frontier model such as ChatGPT-4o (OpenAI et al., 2024), we observe that our model performs considerably better on forward translations (English→Mauritian Creole). However, for reverse translations (Mauritian Creole→English), ChatGPT-4o achieves slightly superior performance on both BLEU and ChrF scores. This asymmetry can likely be attributed to ChatGPT-4o's extensive multilingual training across numerous language pairs, enabling it to leverage cross-lingual priors for improved Mauritian Creole decoding.

Haitian Creole and Mauritian Creole share significant linguistic similarities as French-based creoles developed under similar colonial conditions. They exhibit overlapping vocabulary, grammar, and simplification patterns compared to French (Déprez, 2019), enabling cross-linguistic transfer. Models trained on Haitian Creole can leverage this overlap when processing Mauritian Creole. Given Haitian Creole's much larger training corpus, this likely influences ChatGPT-4o's performance—improving reverse translation but degrading forward translation, as the model tends to apply Haitian grammar to Mauritian output. We observed the same behavior using the Flores-200 Devtest set (see Appendix).

Large language models outperform traditional encoder-decoder architectures in data-scarce scenarios due to their ability to extract linguistic patterns from limited examples. Pre-trained on multilingual corpora, LLMs provide rich contextual representations adaptable to new language pairs with minimal fine-tuning, unlike encoder-decoder models that need substantial parallel data. This advantage is especially important for creole languages, where complex lexifier and substrate influences are better captured by the nuanced knowledge in large-scale pre-trained models.

6 Conclusions

In this paper, we present a new state-of-the-art model for English–Mauritian Creole translation, along with several novel datasets used for training and evaluation. These include: (1) the Flores-200 Dev and Devtest sets, (2) synthetically generated data, (3) a test set from the LALIT newspaper, and (4) community-sourced parallel bitext. This work provides a strong baseline for future model development, which can be improved by collecting more high-quality parallel data. Our results with limited data suggest that incremental augmentation will boost performance, supporting sustainable Mauritian Creole MT development. Additionally, our model can generate high-quality synthetic translations, enabling continual learning through iterative data generation and refinement.

Several promising directions emerge for future research. The inclusion of French parallel bitext represents a particularly valuable avenue, given Mauritian Creole's French lexifier heritage and the abundance of high-quality French-English parallel corpora that could enhance transfer learning effectiveness (Robinson et al., 2023). Incorporating pivot languages; intermediate languages that share linguistic features with both English and Mauritian Creole, could provide additional pathways for cross-lingual knowledge transfer and improved translation quality.

Finally, the systematic generation of synthetic training data through back-translation, paraphrasing, and multilingual data augmentation techniques offers scalable approaches to address the persistent data scarcity challenges that characterize creole language processing. These future developments, building upon the foundation established in this work, promise to advance Mauritian Creole machine translation toward broader practical applicability and community benefit.

Limitations

Our work has several important limitations that should be acknowledged. First, we train our final model on both the FLORES-200 dev and devtest splits, which raises potential evaluation concerns regarding data contamination. While we mitigate this through the use of an independent hold-out evaluation set sourced from LALIT newspaper, the limited size of available high-quality parallel data necessitated this approach to maximize training effectiveness.

Second, our evaluation is constrained by the relatively small size of our test sets (100-1,000 sentences), which may limit the statistical significance and generalizability of our results. The scarcity of Mauritian Creole digital resources inherently constrains the scale of evaluation possible for this language pair.

Finally, our synthetic data generation approach, while innovative, relies on a single large language model (Claude Sonnet 4) and may introduce systematic biases or artifacts that could affect model

performance. The quality and cultural authenticity of synthetically generated Mauritian Creole content, while supplemented with expert knowledge, may not fully capture the nuanced variations present in natural language use.

Acknowledgments

Special thanks to Aishani Rajarai who helped me create and review the high quality parallel bitext datasets. We'd like to also thank Professor David Ifeoluwa Adelani from the Masakhane Community for his guidance.

References

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, and 1 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070. Association for Computational Linguistics.

Anthropic. 2024. Claude sonnet 4. Artificial Intelligence Model. Large language model, version as of June 2024.

Arnaud Carpooran. 2005. *Gramer Kreol Morisien*. Editions Bartholdi, Mauritius. First comprehensive grammar of Mauritian Creole in the language itself.

Raj Dabre and Aneerav Sukhoo. 2022. KreolMorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Michel DeGraff. 2001. On the origin of creoles: A cartesian critique of neo-darwinian linguistics. *Linguistic Typology*, 5(2/3):213–310.

Michel DeGraff. 2005. Linguists' most dangerous myth: The fallacy of creole exceptionalism. *Language in Society*, 34(4):533–591.

Viviane Déprez. 2019. Plurality and definiteness in mauritian and haitian creoles. *Journal of Pidgin and Creole Languages*, 34(2).

Thomas Hylland Eriksen. 2007. Creolization in anthropological theory and in mauritius. In Charles Stewart, editor, *Creolization: History, Ethnography, Theory*, pages 153–177. Left Coast Press, Walnut Creek, CA.

Marcell Fekete, Nathaniel Romney Robinson, Ernests Lavrinovics, Djeride Jean-Baptiste, Raj Dabre, Johannes Bjerva, and Heather Lent. 2025. Limited-resource adapters are regularizers, not linguists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 222–237, Vienna, Austria. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *Preprint*, arXiv:1904.09751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Silvia Kouwenberg and John Victor Singler. 2009. *The Handbook of Pidgin and Creole Studies*. John Wiley & Sons.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, and 1 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. *arXiv preprint arXiv:1808.04189*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Denis Piat. 1999. *Sur la Route des Épices: L'Île Maurice*. Les Editions du Pacifique.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, and 1 others. 2024. Kreyòl-mt: Building mt for latin american, caribbean and colonial african creole languages. *Preprint*, arXiv:2405.05376.

Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Raj Dabre, Kenton Murray, Andre Coy, and Heather Lent. 2025. Findings of the first shared task for creole language machine translation at wmt25. In *Proceedings of the Tenth Conference on Machine Translation*.

Nathaniel Romney Robinson, Matthew Dean Stutzman, Stephen D. Richardson, and David R Mortensen. 2023. African substrates rather than european lexifiers to augment african-diaspora creole translation. In *4th Workshop on African Natural Language Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. *Preprint*, arXiv:1511.06709. [link].

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. *Preprint*, arXiv:1604.02201.

A Training and Evaluation with Flores-200

In this section, we will evaluate the model's performance on FLORES-200 devtest across three modalities: (i) monolingual only, (ii) monolingual + training data, and (iii) monolingual + training + FLORES-200 dev LoRA finetune. We evaluate both translation directions to quantify the additional performance lift from the FLORES dev finetune and assess the validity of using the FLORES-200 devtest set as an evaluation dataset. The LoRA finetune was performed in both translation directions.

Model / Setup	BLEU	CHRF
EN → MFE		
Kreyòl-MT	17.28	49.07
ChatGPT-4o	17.48	48.40
Mono Only	11.94	39.78
Mono + Train	25.76	55.71
Mono + Train + LoRA	**26.83**	**57.68**
MFE → EN		
Kreyòl-MT	28.31	57.29
ChatGPT-4o	**43.08**	**68.76**
Mono Only	33.80	60.88
Mono + Train	40.75	66.65
Mono + Train + LoRA	41.79	67.73

Table 3: BLEU and ChrF scores for different model configurations in EN ↔ MFE translation.

B Dataset Specifics

The training data consisted of KreolMorisienMT (21,810 sentences), KreyolMT (19,149 sentences), Flores Dev/Devtest (2,009 sentences), 300 community-sourced sentences, and synthetic data generated by Claude Sonnet 4. We created parallel bitext datasets (MMLU, QA, and Topic Classification totaling 3,676 sentences) and conversational prompts (1,225 grammar-specific sentences from parsed Mauritian Creole grammar books). For parallel bitext, we used a dual prompt strategy: one prompt asking the model to translate between English and Mauritian Creole, and another presenting questions directly in Mauritian Creole with options and answers in Mauritian Creole.

For monolingual pretraining we only used the monolingual dataset provided by KreolMorisienMT. For the training stage we use both parallel datasets from Kreyol-MT, KreolMorisienMT

and the community sourced bitext. For the fine-tuning stage, we use the flores dev and devtest datasets.

JHU WMT 2025 CreoleMT System Description: Data for Belizean Kriol and French Guianese Creole MT

Nathaniel R. Robinson
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD, USA
nrobin38@jhu.edu

Abstract

This document details the Johns Hopkins University's submission to the 2025 WMT Shared Task for Creole Language Translation. We submitted exclusively to the data subtask, contributing machine translation bitext corpora for Belizean Kriol with English translations, and French Guianese Creole with French translations. These datasets contain 5,530 and 1,671 parallel lines of text, respectively, thus amounting to an 2,300% increase in publicly available lines of bitext for Belizean Creole with English, and an 370% such increase for French Guianese Creole with French. Experiments demonstrate genre-dependent improvements on our proposed test sets when the relevant state-of-the-art model is fine-tuned on our proposed train sets, with improvements across genres of up to 33.3 chrF++.

1 Introduction

The vast majority of countries and territories throughout Central and South American are hispanophone. Two notable exceptions to this trend are anglophone Belize and francophone French Guiana, pictured in Figure 1. These are both home to a duplicity of Creole languages that exist alongside English, French, and other regional languages. Many thousands of Belize residents speak Belizean Kriol and Garifuna. And French Guiana is home to French Guianese Creole, Saramaccan, and Ndyuka (Robinson et al., 2024).

According to the Statistical Institute of Belize' 2022 Population and Housing Census, Belizean Creole is spoken by 181k people in Belize (or 49% of the overall population), making it the third-most-spoken language of the country, after English (at 278k, or 75.5% of the population) and Spanish (at 199k, or 54%).

French Guiana is home to 292k people[1] and a large diversity of both Creole languages (French Guianese Creole, Haitian, French Antillean Creole, Ndyuka, Saramaccan, Sranan Tongo, and Guyanese Creolese); European languages (French, Spanish, English, Dutch, and Portuguese); Amerindian languages (Palikur, Teko, Wayampi, Wayana, and Arawak); and East Asian languages (Hmong, Cantonese, Hakka Chinese) (Léglise, 2013). According to Léglise,[2] French Guianese Creole is the mother tongue of about a third of the population. Ndyuka, an English-related Creole language, and its dialects Aluku and Paramaccan (Hammarström et al., 2023) are also spoken by roughly a third of the population. Multiple of the languages Léglise listed are spoken by immigrant communities: Haitian by 10-20% of the population, French as a mother tongue by roughly 10%, Brazilian Portuguese by 5-10%, and French Antillean Creole languages by roughly 5%. Léglise estimates that less than 5% speak indigenous Amerindian languages.

Figure 1: The greater Caribbean area with Belize and French Guiana indicated in red, made with https://www.mapchart.net

Because English and French are prominent in Belize and French Guiana, respectively, and because these European languages are used by populations that do not speak Belizean Kriol or French Guianese Creole in Belize and French Guiana, we justify the choice of these languages for translation

[1] Per the French Institut national de la statistique et des études économiques

[2] See https://leglise.cnrs.fr/?lang=en.

Proceedings of the Tenth Conference on Machine Translation, pages 1191–1197
November 8-9, 2025

into and out of our Creole languages of focus.

In this work, we contribute:

- A bitext with 5,530 Belizean Kriol and English translations from a Belizean textbook, *Kriol-Inglish Dikshineri/English-Kriol Dictionary* (Herrera et al., 2009)
- A bitext with 879 French Guianese Creole and French translations from a web-sourced Bible text
- A bitext with 792 French Guianese Creole and French translations from a collection of French Guianese fables sourced by the French national library
- State-of-the-art MT improvements across the genres of our newly curated datasets, per our own test sets

2 Related Work

Very little previous work has been published on machine translation (MT) of Belizean Kriol and French Guianese Creole. CreoleVal (Lent et al., 2024), a project introducing a multilingual machine translation model for 28 Creole languages, was the first published work on Belizean Kriol machine translation. Like in this work, Lent et al. focused on translation between Belizean Kriol and English. They put together a Belizean Kriol bitext with English, including 12,085 lines of training data. Using these data, they trained the CreoleM2M model,[3] which acheives 44.4 chrF (Popović, 2015) for Belizean Kriol to English translation and 46.3 chrF for English to Belizean Kriol (on their own test set). However, none of the CreoleVal data were publicly released to the research community.

The other publication to benchmark Belizean Kriol translation was Kreyòl-MT (Robinson et al., 2024), a project that produced a dataset and model for translation of 41 Creole languages, primarily of the Colonial African diaspora. The Kreyòl-MT model[4] supports translation in 172 language directions—all of which include one Creole language and one other language, usually English or French. The Kreyòl-MT model achieves 53.3 chrF on the Kreyòl-MT test set for Belizean Kriol to English translation, and 83.3 chrF for English to Belizean Kriol (on their own private test set). The same model achieves very similar scores on the CreoleVal (Lent et al., 2024) test set: 53.3 chrF into English, and 83.5 into Belizean Kriol, establishing Kreyòl-MT as state-of-the-art for Belizean Kriol translation with English. The dataset used to develop Kreyòl-MT contains 31,002 total lines of bitext for Belizean Kriol aligned with any other language, but only 229 of these lines (all aligned with English translations) were released to the research community in the Kreyòl-MT public dataset.[5]

Kreyòl-MT is also the only published work to address French Guianese Creole MT. On its own test set, Kreyòl-MT achieves 71.9 chrF translating French Guianese Creole into French and 62.4 translating French into French Guianese Creole. This was accomplished using only 292 lines of bitext for training data, indicating that these high scores may be a result of over-fitting to the training domain. This entire bitext was released publicly, along with 50 lines of bitext for testing and 50 for development/validation.

3 Task Description

The first shared task for Creole language machine translation (CreoleMT) for the Tenth Conference on Machine Translation (WMT25) (Robinson et al., 2025) was announced to further advance the state of machine translation for Creole languages. The shared task is comprised of two subtasks: one for systems and one for data. We participate only in the data subtask, which solicits new bitext data for Creole language MT.

As part of the set up for the systems subtask of the shared task, the task organizers established the public Kreyòl-MT dataset as the primary source of training data for the constrained submission track. To make the evaluation fair, they named the Kreyòl-MT **pubtrain** model—an mBART (Liu et al., 2020) initialization developed only with the publicly released Kreyòl-MT train and development sets and hosted at https://huggingface.co/jhu-clsp/kreyol-mt-pubtrain with the name **kreyol-mt-pubtrain**—as the baseline model for benchmarking. (The flagship Kreyòl-MT model was trained on some proprietary data not available for public release and hence would not be a fair baseline.) Since the datasets we curate in this work are publicly releasable, we compare their utility to that of other publicly available data and also adopt **kreyol-mt-pubtrain** as our primary base-

[3] https://huggingface.co/prajdabre/CreoleM2M

[4] https://huggingface.co/jhu-clsp/kreyol-mt

[5] https://huggingface.co/datasets/jhu-clsp/kreyol-mt

line model for MT experiments.

4 Dataset Creation

We detail our creation approaches for all datasets.

4.1 Belizean Kriol dictionary

Our Belizean Kriol-English bitext is taken entirely from aligned sentences in *Kriol-Inglish Dikshineri/English-Kriol Dictionary* (Herrera et al., 2009). The sentences were extracted from a PDF rendering of the book using the PyPDF2 library[6] and software that will be released before publication of this work.

The scraping script used regular expressions to extract all complete sentences used in examples with translations. With the assumption that sentences would alternate between Belizean Kriol and English, we used this software to label every odd sentence as "English" and every even sentence as "Belizean Kriol." As expected, this method of labeling led to a large number of alignment errors.

We corrected these alignment errors via a semi-automated approach. First we trained a decision tree classifier using `scikit-learn`[7] for language identification (LID), to distinguish Belizean Kriol and English sentences. We first used both source and target sides of the Kreyòl-MT public train set for Belizean Kriol-English MT to fit a simple word-based sentence embedder. For 100-dimensional embeddings, the embedder assembled the 100 most common words in the combined source-target corpus. Each sentence is then assigned a binary 100-dimensional vector where the element at each position k indicates the presence or absence of the kth most common word. We then used the same data (the Kreyòl-MT Belizean Kriol-English train bitext) to fit the decision tree, embedding each sentence with our embedder. We used the corresponding Kreyòl-MT validation/development set to evaluate our decision tree, which guided our decision to use 300-dimensional vectors instead of 100-dimensional, and our decision to use a decision tree rather than logistic regression or a random forest. (The decision tree fit with 100-dimensional vectors achieved the highest validation accuracy.)

Next we set up a system where one sentence collection (odd sentences) and the other (even sentences) are initially labeled arbitrarily as "English" and "Belizean Kriol," respectively. Human intervention is requested only whenever the LID classifier does not label both sentences as their expected languages, during which intervention the human can make any adjustments needed, including switching which language corresponds to which sentence set by default.

By this process we achieved a fully aligned bitext of 5,530 lines. Since the previous largest publicly available bitext for Belizean Kriol-English tranaslation was size 240 (Robinson et al., 2024), this amounts to a 2,304% (or 24-fold) increase in publicly available data for the language pair.

The human guiding the alignment process, while not a speaker of Belizean Kriol, speaks English natively and has high familiarity and professional experience with Creole languages. This rendered the alignment guidance easy, and we believe its quality to be high. Table 1 displays a perfectly random sampling of five aligned sentences from the bitext. Note that even non-speakers of Belizean Kriol can verify the validity of these alignments, due to the linguistic proximity of these languages.

4.2 French Guianese YouVersion Bible

The French Guianese Creole biblical data we extracted from the online YouVersion Bible[8] using the BeautifulSoup Python library.[9] The site includes one French Guianese Creole translation of the Gospel of John.[10] We scraped this and aligned it with the corresponding texts in a French Bible translation, the New Geneva Edition, on the same site,[11] resulting in 879 aligned sentences. Since the previous largest public translation dataset for French Guianese Creole-French translation contained 447 aligned sentences (Robinson et al., 2024), this amounts to a 197% increase in publicly available data.

Similar to Table 1, Table 2 displays a perfectly random sampling of five aligned sentences from the bitext.

4.3 French Guianese Gallica fables

We sourced our final bitext from the collection *Introduction à l'histoire de Cayenne: suivie d'un recueil de contes, fables et chansons en créole avec traduction en regard, notes et commentaires* (de Saint-Quentin and de Saint-Quentin, 1872).

[6] https://pypdf2.readthedocs.io/en/3.x
[7] https://scikit-learn.org
[8] https://www.bible.com
[9] https://tedboy.github.io/bs4_doc
[10] https://www.bible.com/bible/2963/JHN.1.GCR07
[11] https://www.bible.com/bible/106/JHN.1.NEG79

English	Belizean Kriol
By the time John was twelve, he was already tall like his dad.	Bai di taim Jan twelv, ih mi don taal laik ih pa.
I'm missing my watch from off the bureau; somebody must have stolen it.	Ah di misn mi wach aaf a mi byooro; sohnbadi mos a teef it.
The woman helped the thieves to escape.	Di uman mi help di teef dehn fi eskayp.
The man took a long time to cut the yard because his machete was dull.	Di man tek lang fi chap di yaad, kaa ih masheet mi dol.
We went to Betty's birthday party last night.	Wi mi gaahn da Beti bertday paati laas nait.

Table 1: Five randomly sampled aligned English and Belizean Kriol translations

French	French Guianese Creole
C'est ici le pain qui descend du ciel, afin que celui qui en mange ne meure point.	Mé dipen ki désann di syèl-a, sala ka manjé li péké mouri.
Le Père aime le Fils, et il a remis toutes choses entre ses mains.	Papa Bondjé kontan so Pitit, é li bay li tout pouvwè ansou tout bagaj.
Pourquoi m'interroges-tu? Interroge sur ce que je leur ai dit ceux qui m'ont entendu; voici, ceux-là savent ce que j'ai dit.	Poukisa to ka kèksyoné mo? Kèksyoné moun-yan ki kouté mo-a, sa-ya, yé byen savé sa mo di yé."
Si je n'étais pas venu et que je ne leur aie point parlé, ils n'auraient pas de péché; mais maintenant ils n'ont aucune excuse de leur péché.	Si mo pa té vini é si mo pa té palé pou yé, yé pa té ké koupab di péché. Mé anprézan, yé pa ganyen pyès èskiz pou yé péché.
Jésus leur répondit: J'ai fait une œuvre, et vous en êtes tous étonnés.	Jézi réponn yé: "Mo fè roun sèl kichoz, é zòt èstébékwé.

Table 2: Five randomly sampled aligned French and French Guianese Creole biblical translations

This digitized book contains French Guianese fables written in French Guianese Creole with French translations. The Gallica website[12] from which we sourced this text contained eight such fables, organized into a total of 48 sections. Once again using BeautifulSoup for web-scraping, we extracted the sentences and aligned the text sections automatically. Sentence alignment was then performed manually, along with a substantial amount of error correction; many of the text segments on the web-page had what appeared to be OCR-induced noise.[13] As a disclaimer, the author who performed this manual alignment and cleaning is not proficient in French Guianese Creole. However, he is proficient in both French and another French-related Creole language (Haitian), and he has both linguistic training and experience in bitext development. Sentence alignment and cleaning decisions were mostly made clear by cognates and context. The changes to the originally web-scraped document are publicly available for perusal on GitHub.[14]

Similar to Tables 1 and 2, Table 3 displays five randomly sampled aligned and cleaned sentence pairs from the fables dataset.

5 Dataset Details

Our datasets consist of one genre each, educational material in Belizean Kriol and English, and Biblical text and literature in French Guianese Creole and French. Both datasets we shuffled randomly and then split into train, dev, and test sets with an 80-10-10 split ratio. See the shared task findings (Robinson et al., 2025) for data cards of our submitted datasets.

5.1 Fulfillment of task requirements

We now detail how our datasets fulfill each of the data requirements set forth for the shared task.[15]

1. **Translations in datasets must be completely conducted by native or proficient speakers of both languages.** The translations in our Belizean Kriol-English bitext come from *Kriol-Inglish Dikshineri/English-Kriol Dictionary* (Herrera et al., 2009), a publication authorized by the Belize Ministry of Education. The French Guianese Creole-French sentences come from Bible translations and fable

[12] https://gallica.bnf.fr/ark:/12148/bpt6k82939m/texteBrut, accessed August 2025

[13] "Typos" caused by errors in an optical character recognition system that may have digitized the text from an original print source

[14] https://github.com/n8rob/creolemt_wmt25_jhu_submission/pull/1/files

[15] https://www2.statmt.org/wmt25/creole-mt.html

French	French Guianese Creole
Il resta ainsi longtemps,	Li rété bon moso konsa
Si vous voulez vous laver dedans,	Si zôt oulé lavé landan,
Il en restait encore une assez grande quantité;	Bon moso té rété enko
sans pouvoir en prendre un seul pour mon souper.	san mo pa pouvé kienbé oun pou mo sonpé.
J'en retirerai un peu avec mes dents »	M'a tire moso ké mo dan »

Table 3: Five randomly sampled cleaned and aligned French and French Guianese Creole fable translations

translations organized by the Bibliothèque nationale de France. The reputability of these sources convince us that such translations were performed by competent translators.

2. **Participants must demonstrate convincingly that one language in each submitted bitext is considered a Creole language.** Both Belizean Kriol and French Guianese Creole are universally considered Creole languages as attested by their inclusion in APICS (Michaelis et al., 2013), among other sources (McWhorter, 2005; DeCamp, 1968).

3. **If submitting a test set, participants must use it to evaluate performance of an MT model and provide compelling evidence that performance aligns with conventional wisdom.** We include results to address this point, for all three of our submitted datasets, in §6.

6 Experimental Results

Here we detail experiments that minimally show the utility of our datasets, as requested by the shared task organizers. We show both SpBLEU (Papineni et al., 2002; Goyal et al., 2022) and chrF++ scores (Popović, 2015, 2017) for the **`kreyol-mt-pubtrain`** model both out of the box, and after fine-tuning on our train sets (with our dev sets used to determine time of early stopping). We trained with a learning rate of $2 * 10^{-5}$, a batch size of 4, a weight decay of 0.01, and a maximum of 10 training epochs. We implemented early stopping with a stopping patience of 2 and a stopping threshold of 0.01. We fine-tuned each model on both translation directions for the language pair in question, and with early stopping we ended up finetuning for Belizean Kriol-English translation for 13,272 training steps, and for French Guianese Creole-French translation for 4,683 training steps.

	eval	score	OOB	FT
`bzj→eng`	*Dict.*	SpBLEU	17.0	**52.1**
		chrF++	37.9	**66.8**
`eng→bzj`	*Dict.*	SpBLEU	8.46	**48.1**
		chrF++	24.8	**58.1**
`gcr→fra`	*Bible*	SpBLEU	18.1	**35.3**
		chrF++	39.1	**51.7**
	Fable	SpBLEU	36.6	**41.3**
		chrF++	51.5	**54.7**
`fra→gcr`	*Bible*	SpBLEU	6.86	**25.4**
		chrF++	20.5	**34.8**
	Fable	SpBLEU	34.0	**37.9**
		chrF++	44.8	**48.6**

Table 4: MT scores for **`kreyol-mt-pubtrain`** out-of-the-box (OOB) and fine-tuned on our novel train sets (FT). Both SpBLEU and chrF++ scores are computed on our proposed test sets. Better results are **bold**.

Table 4 displays our MT performance scores. Fine-tuning on our training data significantly improves performance on our test sets in every circumstance. Some improvements are large, but their magnitude depends on test set language and genre. Out of the Creole language and into the Creole language, respectively, chrF++ improves by 28.9 and 33.3 on the Belizean dictionary test set, by 12.6 and 14.3 on the Guianese Bible test set, and by 3.2 and 3.8 on the Guianese fables test set. Note that Table 4 uses ISO 639-3 codes to abbreviate language names.[16]

7 Conclusion

We introduce three new datasets for Creole language translation of two language pairs:

1. **A bitext with 5,530 aligned translations of the educational genre in Belizean Kriol and English**, resulting in a 2,304% increase

[16] `bzj` = Belizean Kriol; `eng` = English; `gcr` = French Guianese Creole; `fra` = French

of publicly available bitext data for the language pair, and improvements of 28.9 chrF++ into-English and 33.3 chrF++ out-of-English over the state-of-the-art MT model trained on publicly available Kreyòl-MT data on our in-genre test set

2. **Two bitexts with combined 1,671 aligned translations in French Guianese Creole and French: 879 biblical sentences and 792 lines of Guianese folktales**, resulting in a 370% increase of publicly available bitext data for the language pair; improvements of 12.6 chrF++ into-French and 14.3 chrF++ out-of-French over the state-of-the-art MT model trained with publicly available data on our Biblical test set, and like improvements of 3.2 into-French and 3.8 out-of-French on our fables test set

We hope these incremental contributions may forward research in MT for Creole languages.

Limitations

Note that due to genre constraints, the utility of these datasets may be restricted to certain domains. For instance, the French Guianese Creole-French bitext we curated may be useful for assistance in further Bible translation, but may have limited utility beyond that.

Additionally, our compiling of the contributed datasets was constrained by time. We hope in future shared tasks to contribute even more data as we have the time to collect it. We also note that even in conducting manual cleaning and alignment of the datsets we submit here, cleaning and alignment errors can still be found in some places of the datasets. Though it may not be possible for these low-resource datasets to be completely noise-free, we hope future shared tasks will afford us the time to make extra passes and clean more thoroughly.

References

Alfred de Saint-Quentin and Auguste de Saint-Quentin. 1872. *Introduction à l'histoire de Cayenne: suivie d'un recueil de contes, fables et chansons en créole avec traduction en regard, notes et commentaires*. J. Marchand.

David DeCamp. 1968. The field of creole language studies. *Latin American Research Review*, 3(3):25–46.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Harald Hammarström, Sebastian Nordhoff, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. Glottolog. Online database.

Yvette Herrera, Myrna Manzanares, Silvaana Udz, Cynthia Crosbie, and Ken Decker. 2009. Kriol-inglish dikshineri/english-kriol dictionary. *Belize Kriol Project.–Belmopan, Belize*, 465.

Isabelle Léglise. 2013. *Multilinguisme, variation, contact. Des pratiques langagières sur le terrain à l'analyse de corpus hétérogènes*. Accreditation to supervise research, Institut National des Langues et Civilisations Orientales- INALCO PARIS - LANGUES O'.

Heather Lent, Kushal Tatariya, Raj Dabre, Yiyi Chen, Marcell Fekete, Esther Ploeger, Li Zhou, Ruth-Ann Armstrong, Abee Eijansantos, Catriona Malau, Hans Erik Heje, Ernests Lavrinovics, Diptesh Kanojia, Paul Belony, Marcel Bollmann, Loïc Grobol, Miryam de Lhoneux, Daniel Hershcovich, Michel DeGraff, and 2 others. 2024. CreoleVal: Multilingual multitask benchmarks for creoles. *Transactions of the Association for Computational Linguistics*, 12:950–978.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

John H McWhorter. 2005. *Defining creole*. Oxford University Press.

Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. 2013. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Stutzman, Bismarck Odoom, Sanjeev Khudanpur, Stephen Richardson, and Kenton Murray. 2024. Kreyòl-MT: Building MT for Latin American, Caribbean and colonial African creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110, Mexico City, Mexico. Association for Computational Linguistics.

Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Raj Dabre, Kenton Murray, Andre Coy, and Heather Lent. 2025. Findings of the first shared task for creole language machine translation at wmt25. In *Proceedings of the Tenth Conference on Machine Translation*.

WMT 2025 CreoleMT Systems Description : Martinican Creole and French

Ludovic Mompelat
Department of Modern Languages
University of Miami
Miami, FL, USA
lvm861@miami.edu

Abstract

This paper describes our submissions to the constrained subtask of the WMT25 Creole Machine Translation shared task. We participated with a bidirectional Martinican Creole ↔ French system. Our work explores training-time strategies tailored for low-resource MT, including LoRA fine-tuning, curriculum sampling, gradual unfreezing, and multitask learning. We report competitive results against the baseline on both translation directions.

1 Introduction

The WMT25 Creole MT shared task is the first shared task focused on machine translation involving Creole languages. As part of the WMT25 Creole MT shared task (Robinson et al., 2025), we submitted constrained systems for Martinican Creole ↔ French. This track explores how to improve Creole MT under limited resource conditions. Participants were restricted to using only the official training data provided by the organizers and were allowed to initialize their models from the baseline `kreyol-mt-pubtrain` model on HuggingFace.

Our submission focused on the Martinican Creole ↔ French language pair (ISO: `mart1259-fra`). We trained and evaluated a bidirectional system for both `mart1259`→`fra` and `fra`→`mart1259` directions.

Motivation Our participation stems from ongoing work on NLP tools for Creole languages, including Martinican Creole and Haitian Creole. Despite progress in NLP, Creole languages remain underrepresented and under-resourced, making them ideal candidates for constrained MT settings. We aimed to investigate how far one can push model performance without access to expanded corpora by instead exploring training strategies and fine-tuning configurations tailored to low-resource language modeling.

This work is part of a broader effort to develop NLP tools for Creole researchers, educators, and communities. In particular, we aim to build usable MT systems, including domain-specific applications (e.g., medical MT for Haitian Creole). We explore various configurations: custom tokenizer, curriculum sampling, LoRA fine-tuning, gradual unfreezing, denoising, label smoothing, weighted BLEU scoring, multitask learning, and mixed training strategies.

2 Task Overview and Constraints

The WMT25 constrained track permits only the use of official datasets provided by the organizers. No additional monolingual or parallel data was allowed for training, tuning, or backtranslation. The baseline system is the publicly available `kreyol-mt-pubtrain` model on HuggingFace.

2.1 Data Used

We used the official datasets for the `mart1259-fra` pair, cleaned and released by the organizers:

- **Dictionnaire créole martiniquais–français** by Raphaël Confiant (2007)[1].
- CAPES corrections for Guadeloupean Creole from the Kapes Kreyol project.

2.2 Baseline Scores

mart1259→fra. On the organizers' blind test set, our system scores **25.31 BLEU** and **49.08 chrF++**, below the baseline (**28.27 BLEU**, **50.43 chrF++**). While LoRA plus curriculum sampling proved stable during development, these settings did not surpass the strong baseline under official evaluation.

fra→mart1259. On the blind test set, our system achieves **25.76 BLEU** and **48.74 chrF++**, close to but slightly below the baseline (**26.49 BLEU**, **48.69 chrF++**). This direction is comparatively tighter,

[1] https://www.potomitan.info/dictionnaire/

Proceedings of the Tenth Conference on Machine Translation, pages 1198–1200
November 8-9, 2025

suggesting our configuration is broadly competitive but does not yield consistent gains over baseline.

3 Data Preprocessing and Setup

We experimented with a 70/30 training/validation split (3,094 / 1,338) to increase the size of the development set for more stable evaluation but reverted back to the original 90/10 split as it yielded slightly better results. Training was conducted separately for each translation direction rather than using concatenated bidirectional data or multitask learning, as this yielded more consistent results in our preliminary experiments given the dataset provided.

3.1 Tokenization and Language Codes

We retained the baseline's SentencePiece model and language code scheme (e.g., `mart1259`, `fra`) to ensure vocabulary alignment. Custom `forced_bos_token_id` was applied during decoding to enforce direction.

3.2 Curriculum Sampling and Difficulty Weighting

We computed difficulty scores for each example and implemented curriculum sampling in early epochs, gradually introducing harder examples. Weighted BLEU was used for dev set evaluation to better align training progress with final performance.

4 Model Architecture and Training

All systems were initialized from the publicly available `jhu-clsp/kreyol-mt-pubtrain` model on HuggingFace. We applied Low-Rank Adaptation (LoRA) for efficient fine-tuning, targeting the attention on `q_proj` and `v_proj`. Our final configuration used rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.1, though we also explored $r \in \{8, 32\}$ and $\alpha \in \{64\}$ in preliminary runs.

Training was performed with a custom `Seq2SeqTrainer` that integrates curriculum sampling (ordering examples by difficulty), gradual unfreezing of encoder and decoder layers, and multitask loss weighting to balance direction-specific performance. We applied label smoothing of 0.1 to improve generalization and used weighted BLEU scoring on the development set as the main model selection criterion.

The final training configuration included a batch size of 32, maximum source and target lengths of 128 tokens, 50 training epochs, a constant learning rate of 2×10^{-5} with 500 warmup steps, and the AdamW optimizer. Evaluation used a beam size of 4 and a maximum of 128 new tokens during generation.

5 Evaluation

5.1 Final Systems

For `mart1259`→`fra`, our final constrained submission fine-tunes the publicly available `jhu-clsp/kreyol-mt-pubtrain` model, an mBART-50 style encoder–decoder Transformer, using only the official training set with the *original* 90/10 split. This direction uses Low-Rank Adaptation (LoRA) and curriculum sampling with difficulty weighting, as these settings yielded the most stable gains over the baseline.

For `fra`→`mart1259`, preliminary experiments showed that a 70/30 training/validation split improved dev set stability, while curriculum sampling provided no measurable benefit. We therefore retained LoRA but removed curriculum sampling for this direction.

In both cases, we retain the original SentencePiece tokenizer and language tags, forcing the target language with `decoder_start_token_id`. Training uses a batch size of 32, label smoothing (0.05), AdamW with a learning rate of 2×10^{-5}, cosine scheduling, and early stopping based on BLEU on the dev set. Generation uses beam search (beam size 4, max length 128 tokens).

Parameter	Value
Batch size	32
Max source/target length	128 / 128
Epochs	50
Learning rate	2e−5
LoRA rank r	16
LoRA α	32
LoRA dropout	0.05
Target modules	q_proj, v_proj
Label smoothing	0.1
Optimizer	AdamW
Warmup steps	500
Scheduler	constant
Beam size (eval)	4
Max new tokens (eval)	80

Table 1: Training and generation hyperparameters for both final constrained submissions.

No additional data or back-translation was

used. We fine-tune the publicly released `jhu-clsp/kreyol-mt-pubtrain` model with parameter-efficient adaptation (LoRA). Decoding uses beam search (beam size 4; max length 128), and we constrain the target language with `decoder_start_token_id`.

We report **BLEU** (Papineni et al., 2002) and **chrF++** (Popović, 2016). These are the *official* blind-test scores provided by the organizers (Robinson et al., 2025).

System	BLEU	chrF++
Baseline (official)	28.27	50.43
Ours (primary, official)	25.31	49.08

Table 2: `mart1259`→`fra` official blind-test results.

On the organizers' blind test set, our system is below the baseline (25.31 vs. 28.27 BLEU; 49.08 vs. 50.43 chrF++). While LoRA with curriculum sampling was stable in development, it did not surpass the strong baseline under official evaluation for this direction.

System	BLEU	chrF++
Baseline (official)	26.49	48.69
Ours (primary, official)	25.76	48.74

Table 3: `fra`→`mart1259` official blind-test results.

For `fra`→`mart1259`, our system is close to but slightly below the baseline in BLEU (25.76 vs. 26.49), with nearly identical chrF++ (48.74 vs. 48.69). The 70/30 split without curriculum sampling remained a stable configuration for this direction, but it did not yield a clear improvement over the baseline on the official evaluation.

Overall, direction-specific tuning—retaining curriculum sampling for `mart1259`→`fra` and removing it for `fra`→`mart1259`—produced systems that are broadly competitive but ultimately *do not surpass* the baseline under the constrained track conditions. In particular, the gap is larger for `mart1259`→`fra` and smaller for the reverse direction. A plausible explanation is the limited size and diversity of the training data: with few examples, systems may be less sensitive to changes in fine-tuning strategy or adapter placement, and estimates of improvements can have high variance. We therefore treat these direction-specific effects as suggestive rather than definitive.

6 Conclusion and Future Work

Under the constrained track, our LoRA-based systems with direction-specific training choices are close to but ultimately below the strong baseline on the organizers' blind test set. The gap is larger for `mart1259`→`fra` and smaller for `fra`→`mart1259`, where BLEU and chrF++ are nearly tied. One plausible factor is data scarcity: with limited and relatively homogeneous training material, core architectural or training-time changes may yield muted or unstable gains, and measurement noise can obscure small effects.

We will (i) explore alternative parameter-efficient adapters and layer targeting paired with data-scaling, (ii) tune curriculum schedules per direction and analyze where they help/hurt, (iii) align preprocessing/normalization and decoding with the organizers' pipeline, (iv) investigate tokenizer and segmentation choices for Martinican Creole, and (v) add robust automatic and human error analyses (e.g., code-switching, diacritics, OOVs). In the unconstrained setting, we also plan to study back-translation and multilingual transfer to further close the gap.

References

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2016. chrf deconstructed: beta parameters and n-gram weights. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 499–504.

Nathaniel R. Robinson, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Raj Dabre, Kenton Murray, Andre Coy, and Heather Lent. 2025. Findings of the first shared task for creole language machine translation at wmt25. In *Proceedings of the Tenth Conference on Machine Translation*.

JU-NLP: Improving Low-Resource Indic Translation System with Efficient LoRA-Based Adaptation

Priyobroto Acharya[1], Haranath Mondal[2], Dipanjan Saha[3], Dipankar Das[4], Sivaji Bandyopadhyay[5]
[1]Dept. of Power Engineering, Jadavpur University, Kolkata, India
[2,3,4,5]Dept. of CSE, Jadavpur University, Kolkata, India
{ priyobrotoacharya98, haranathnlp, sahadipanjan6, dipankar.dipnil2005, sivaji.cse.ju}@gmail.com

Abstract

Low-resource Indic languages such as Assamese, Manipuri, Mizo, and Bodo face persistent challenges in NMT due to limited parallel data, diverse scripts, and complex morphology. We address these issues in the WMT 2025 shared task by introducing a unified multilingual NMT framework that combines rigorous language-specific preprocessing with parameter-efficient adaptation of large-scale models. Our pipeline integrates the NLLB-200 and IndicTrans2 architectures, fine-tuned using LoRA and DoRA, reducing trainable parameters by over 90% without degrading translation quality. A comprehensive preprocessing suite, including Unicode normalization, semantic filtering, transliteration, and noise reduction, ensures high-quality inputs, while script-aware post-processing mitigates evaluation bias from orthographic mismatches. Experiments across English↔Indic directions demonstrate that NLLB-200 achieves superior results for Assamese, Manipuri, and Mizo, whereas IndicTrans2 excels in English↔Bodo. Evaluated using BLEU, chrF, METEOR, ROUGE-L, and TER, our approach yields consistent improvements over baselines, underscoring the effectiveness of combining efficient fine-tuning with linguistically informed preprocessing for low-resource Indic MT.

1 Introduction

Low-resource Indic languages such as Assamese (As), Manipuri (Mni), Mizo (Lus), and Bodo (Brx) pose significant challenges for **N**eural **M**achine **T**ranslation (NMT) due to data scarcity, script diversity, and linguistic complexity, often leading to suboptimal performance (Kunchukuttan, 2020a; Ramesh et al., 2023; Team et al., 2022a). This work aims to address these limitations by developing an efficient, parameter-optimized fine-tuning framework tailored for such underrepresented languages in the WMT 2025 shared task (Pakray et al.).

To address these gaps, we introduce a unified multilingual NMT pipeline tailored for low-resource Indic languages, combining robust preprocessing with parameter-efficient fine-tuning methods. We integrate **N**o **L**anguage **L**eft **B**ehind (NLLB-200) model (Team et al., 2022a) and **IndicTrans**2 (Ramesh et al., 2023) model, fine-tuning them using **Lo**w-**R**ank **A**daptation (LoRA) as proposed by Hu et al. (2021a) and Weight-**D**ecomposed L**o**w-**R**ank **A**daptation (DoRA) as discussed by Zhao et al. (2023) to optimize performance while maintaining computational efficiency. Our preprocessing pipeline includes Unicode normalization, semantic filtering, transliteration (Kunchukuttan, 2020a), and noise reduction, ensuring high-quality input data for training. NLLB-200, with its extensive multilingual coverage, is adapted for En↔As, Mni, and Lus, while IndicTrans2, designed specifically for Indic languages, is fine-tuned for En↔Brx to leverage its architectural strengths in low-data settings. The methodology ensures fair model comparison by maintaining consistent hyperparameters and evaluation settings across all language pairs, with key contributions lying in the combination of efficient fine-tuning, language-specific preprocessing, and script normalization for Indic NMT.

Our contributions include: (1) the first systematic application of LoRA/DoRA to NLLB-200 and IndicTrans2 for low-resource Indic languages, reducing trainable parameters by over 90% without sacrificing translation quality; (2) a novel preprocessing framework addressing script diversity and data noise, critical for morphologically complex languages; and (3) a comprehensive evaluation using BLEU (Papineni et al., 2002a), chrF

Proceedings of the Tenth Conference on Machine Translation, pages 1201–1209
November 8-9, 2025

(Popović, 2015), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and TER (Snover et al., 2006) metrics, demonstrating significant improvements over baseline approaches.

2 Related Work

Early work on translation involving Indic languages predominantly used statistical methods and ad-hoc bilingual corpora. For example, Koehn (2005a) introduced the *Europarl* corpus for SMT, but no comparable large-scale corpus existed for low-resource Indian languages (Kakum et al., 2023). In practice, government and academic groups built phrase-based systems on much smaller data. India's *TDIL* mission developed the *Sampark* and *Anuvadaksha* translation programs by training phrase-based SMT models on limited domain-specific corpora. Similarly, Kunchukuttan and Bhattacharyya (2014) compiled *Sata Anuvadak*, a set of 110 SMT systems across Indian language pairs. These efforts established early benchmarks but exposed severe limitations due to data scarcity and domain mismatch.

With the advent of neural models, encoder–decoder architectures with attention (Bahdanau et al., 2015) and Transformers (Vaswani et al., 2017) became standard. Researchers trained RNN and then Transformer-based NMT systems for English–Hindi and other Indic pairs, often using byte-pair encoding and shared vocabularies. Multilingual and zero-shot strategies (Johnson et al., 2017) enabled parameter sharing across related languages, benefiting extremely low-resource pairs. Shared multilingual models improved translation quality through inductive transfer, as shown in early WMT shared tasks (Pal et al., 2023). Indic-to-Indic multilingual training further enhanced performance in cases of limited parallel data (Pakray et al., 2024).

In recent years, large multilingual pre-trained models have been employed for Indic MT. Models like mBART (Liu et al., 2020) and mT5 (Xue et al., 2021) provide off-the-shelf improvements, even for Indian languages. In parallel, Indic-specific models such as IndicBART (Dabre et al., 2022) and IndicTrans2 (Ramesh et al., 2023) have emerged. These models were trained on carefully normalized Indic corpora and have shown superior performance in low-resource translation. IndicTrans2, in particular, supports translation across all 22 scheduled Indian languages and 462 Indic language pairs, making it one of the most comprehensive Indic MT systems.

More recently, ultra-large multilingual models and efficient fine-tuning methods have influenced this domain. The NLLB-200 model (Team et al., 2022b) introduced a massively multilingual architecture covering 200 languages, with strong performance on low-resource Indic pairs. To adapt such models efficiently, LoRA (Hu et al., 2021b) and DoRA (Zhao et al., 2023) have been proposed, drastically reducing fine-tuning cost while preserving performance. Finally, preprocessing methods such as Unicode normalization, script unification, and transliteration (Kunchukuttan, 2020b) have been shown to significantly enhance translation quality for Indic languages. These developments form the foundation for recent SOTA systems tailored to low-resource Indic MT.

3 Analysis of Dataset

For the machine translation experiments, we utilized the **WMT** 2025 corpus divided into two categories: **Category-**1 (En $\leftrightarrow$ {As, Lus, Mni}) with moderate training data availability, and **Category-**2 (En $\leftrightarrow$ Brx) with limited training data. The following sections detail each language pair's parallel corpus specifications.

Table 1: Parallel sentences dataset statistics for both Category-1 and 2.

Lang Pair	Script	Dataset	Parallel sents
En - As	Bengali	Training	50000
		Validation	2000
		Test	2000
En - Mni	Bengali	Training	21687
		Validation	1000
		Test	1000
En - Lus	Latin	Training	50000
		Validation	1500
		Test	2000
En - Brx	Devanagari	Training	13693
		Validation	1000
		Test	1000

Table 1 summarizes the dataset sizes and scripts used for each language pair. The pairs En-As and En-Lus have the largest training sets ($50k$ sentences each), while the smallest ones are En-Mni and En-Brx ($21,687$ and $13,693$ sentences, respectively). All language pairs are divided into validation and test sets, where En-As and En-Lus have a larger test

set (2,000 sentences each), followed by En-Mni and En-Brx (1,000 sentences each). The scripts are different by language, using Bengali for As and Mni, Latin for Lus, and Devanagari for Brx.

Table 2: Sentence-level statistics for parallel corpora across four Indic language pairs.

Lang Pair	Avg. Sent. Length	Pearson Correlation	Unique Chars
En - As	En: 95.12 As: 91.29	0.7288	En: 137 As: 187
En - Mni	En: 102.79 Mni: 103.70	0.9447	En: 145 Mni: 177
En - Lus	En: 95.81 Lus: 97.73	0.8843	En: 119 Lus: 136
En - Brx	En: 96.07 Brx: 101.77	0.9377	En: 114 Brx: 144

Table 2 shows sentence-level statistics of the parallel corpora and illustrates the observed linguistic differences in the language pairs. The average number of words in an English sentence (En) ranges from 95.12 (En-As) to 102.79 (En-Mni). On the contrary, for target languages, the average number of words in a sentence is nearly the same or slightly longer with Manipuri (Mni) at 103.70 and Bodo (Brx) at 101.77. The Pearson correlation coefficients, which measure the degree of alignment of sentence lengths of English with the target languages, show that En-Mni (0.9447) and En-Brx (0.9377) have almost a perfect linear relationship, indicating highly consistent translation lengths. In contrast, En-As is least correlated (0.7288), meaning sentence lengths vary more across translations. The unique character count further reflects script complexity, with Assamese (As: 187), Manipuri (Mni: 177), Bodo (Brx: 144), and Mizo (Lus: 136). These statistics emphasize the diversity of languages in the data, which impacts translation modeling, especially for languages with rich morphology or weaker sentence-length correlation.

4 Methodology and Implementation Details

4.1 Data Preprocessing

- ***Unicode normalization*** is essential for machine translation in Indic languages because it ensures consistent text representation by converting multiple Unicode forms into a standardized format, improving tokenization, reducing noise, and enhancing alignment in parallel data. We have used `IndicNormalizer`[1] for Indic languages like Assamese and `unicodedata`[2] **N**ormalization **F**orm-**K** **C**anonical Composition (NFKC) normalizer for English language.

- ***Deduplication*** removes duplicate sentence pairs from parallel corpora, maximizing data utility for low-resource Indic machine translation. This is implemented by Python's built-in library `set()`, which removes duplicate sentence pairs from datasets.

- ***Ratio Filtering*** is essential in machine translation to ensure balanced sentence-length pairs by removing extreme mismatches, which could otherwise introduce noise and misalignment during training. Here, the implementation checks if the **word-count ratio** falls within `0.5` to `2.0`, retaining only pairs where the target sentence is neither half nor double the source length, thus preserving linguistically plausible alignments (Koehn, 2005b).

- ***Semantic filtering*** is crucial for Indic language machine translation to remove poorly aligned bilingual pairs that share surface-level similarities but differ in meaning. This is implemented using LaBSE (Feng et al., 2022), a multilingual sentence embedding model trained on 109 languages through a translation ranking objective, which provides language-agnostic representations without requiring task-specific fine-tuning. We apply cosine similarity scoring between sentence embeddings, where pairs scoring below a 0.75 threshold are excluded from training data to preserve semantic integrity. In our setup, we employ LaBSE specifically for English-side filtering to ensure high-quality parallel data alignment.

- ***Length filtering*** is essential for machine translation to exclude excessively long sentences that may exceed model context limits or contain noisy data. This is implemented through a simple character count check (`150` words maximum per sentence) applied uniformly to both source and target texts.

[1] https://github.com/anoopkunchukuttan/indic_nlp_library

[2] https://docs.python.org/3/library/unicodedata.html

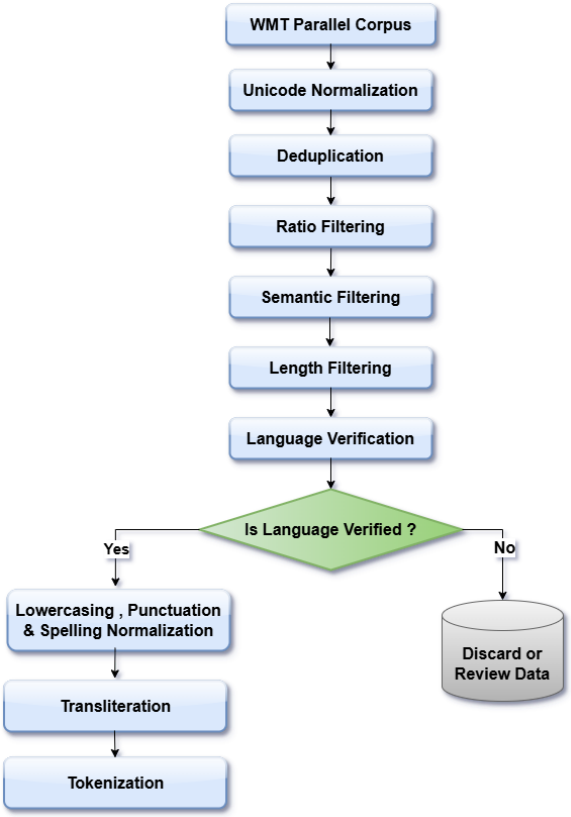


Figure 1: Workflow diagram of proposed data preprocessing pipeline.

- ***Language filtering***: To maintain high-quality, language-specific data for low-resource Indic machine translation, we employ FastText's pretrained language identification model (`ft_model`)(Joulin et al., 2017) to filter out noisy or mixed-language text. The sentences that are not confidently predicted as the target language are removed from the training corpus. Suspicious samples are retained for manual review to either: (1) salvage valuable translation pairs, or (2) analyze common noise patterns that could inform future data collection (Caswell et al., 2019).

- ***Text normalization***: We perform lowercasing, punctuation standardization, and spelling normalization (handling common orthographic variants) to reduce vocabulary sparsity. Aggressive noise removal eliminates HTML tags, non-linguistic symbols, and irregular whitespace, particularly crucial for noisy user-generated content in low-resource languages like Assamese.

- ***Transliteration*** is essential for handling named entities and rare words in low-resource Indic language machine translation. We implement a selective transliteration pipeline using `spaCy`[3] for tokenization and **N**amed **E**ntity **R**ecognition(NER), identifying words with frequency less than or equal to 2 or labeled as named entities. These words are transliterated from English to Indic scripts such as Assamese, Manipuri, and Mizo using the `IndicTransliteration` library[4], via the **H**arvard-**K**yoto (HK) scheme. This preserves phonetic structure and improves source-target alignment, enhancing overall translation quality.

- ***Tokenization*** splits text into subword units, crucial for handling morphologically rich Indic languages by addressing vocabulary sparsity and **O**ut-**o**f-**V**ocabulary (OOV) issues. For Assamese, Manipuri, and Mizo, we use Facebook's `NLLB-200-3.3B` tokenizer with a forced **B**eginning **O**f **S**equence (BOS) token for target language specification. For Bodo, we employ AI4Bharat's `Indictrans2` tokenizer, which supports multiple Indic languages via subword segmentation. Both tokenizers ensure compatibility with their respective Seq2Seq models by setting padding tokens dynamically.

4.2 Approach

This work utilizes the **WMT dataset** provided by the organizers. Consistent with established methodology for low-resource NMT, the data underwent preprocessing (detailed in Section 3) before model input to optimize translation quality for the target Indic languages. Given the focus on low-resource languages, specifically **Assamese, Manipuri, Mizo, and Bodo**, the model training pipeline is designed to leverage existing multilingual capabilities. In this study, two **s**tate-**o**f-**t**he-**a**rt (SOTA) open-source multilingual NMT models with pre-trained Indic language support are evaluated. Both models are subsequently fine-tuned on the preprocessed WMT dataset using LORA for parameter efficiency. Model selection is determined by comparative evaluation across standard automatic metrics: *BLEU*, *chrF*, *METEOR*, *ROUGE-L*, and *TER*.

[3] https://github.com/explosion/spaCy

[4] https://github.com/indic-transliteration/indic_transliteration

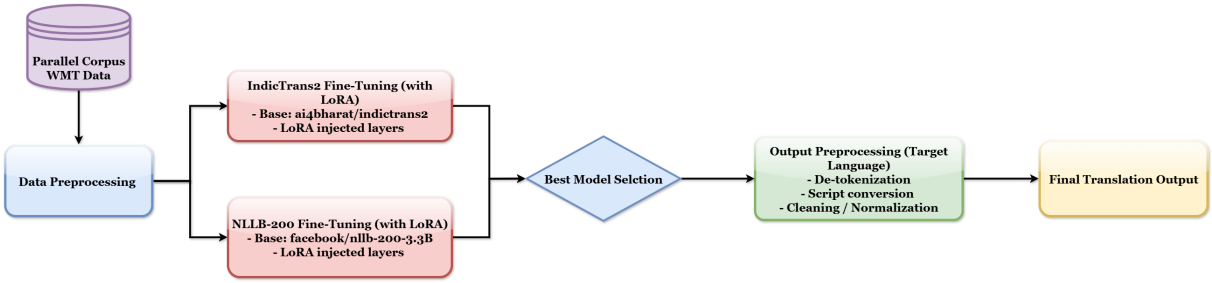


Figure 2: Bird's Eye View of the Proposed Approach

The NLLB-200 model, developed by Meta AI, is a 3.3 billion-parameter multilingual sequence-to-sequence transformer that supports translation across 200 languages, including many low-resource ones, achieving SOTA performance. To fine-tune this model efficiently while preserving its generalization capabilities, we employ **P**arameter-**E**fficient **F**ine-**T**uning (PEFT) as discussed by Xu et al. (2023) via LoRA. This approach avoids full-model fine-tuning by instead injecting trainable low-rank matrices into the transformer's attention layers, drastically reducing the number of trainable parameters while maintaining strong downstream task performance. The LoRA configuration is applied to the query, key, value, and output projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) of the NLLB-200 model. We set the rank (r) of the low-rank matrices to 64, with a scaling factor `lora_alpha` (α) of 128 to balance adaptation strength. A dropout rate of 0.1 is applied to the LoRA layers for regularization, and no additional bias terms are introduced. The model is then converted into a PEFT model, and all trainable parameters are logged before transferring the model to a CUDA-enabled 2× T4 Tesla GPU for accelerated training.

To handle variable-length sequences efficiently, we use a data collator specifically designed for sequence-to-sequence tasks. This collator dynamically pads input sequences to the longest length in each batch while ensuring padding aligns to multiples of 8 for optimal hardware utilization (Wolf et al., 2020). Label padding tokens (set to -100) are masked to exclude them from loss computation during training (Lewis et al., 2020). The training process leverages mixed-precision (FP16) arithmetic via the `Seq2SeqTrainer` from the Hugging Face Transformers library (Wolf et al., 2020). We employ a global batch size of 8, achieved through a per-device batch size of 4 and 2 gradient accumulation steps, balancing training stability (Micikevicius et al., 2018).

The optimization process uses AdamW with fused CUDA kernels (`adamw_torch_fused`), configured with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$ (Loshchilov and Hutter, 2019). The learning rate follows a cosine decay schedule, starting from 3×10^{-5} with 1000 warmup steps to ensure stable early training (Loshchilov and Hutter, 2016). Model checkpoints are saved at the end of each epoch, with the best model selected based on BLEU score (higher is better) (Papineni et al., 2002b). To improve evaluation efficiency, the trainer is configured to generate predictions during validation, enabling direct computation of translation metrics. To optimize memory efficiency, we disable caching (`model.config.use_cache = False`), enabling gradient checkpointing at the cost of modest recomputation (Chen et al., 2016). The complete training system integrates our LoRA-adapted NLLB-200 model with dynamic batching and automated evaluation, maintaining multilingual capabilities while specializing for target domains. This approach enables efficient adaptation of the $3.3B$-parameter model, particularly valuable for low-resource languages where data efficiency is critical (Team et al., 2022a). The implementation demonstrates practical fine-tuning of massive multilingual models within resource constraints, balancing computational feasibility with translation quality.

On the other hand, the IndicTrans2, another state-of-the-art multilingual NMT model developed by AI4Bharat, supports translation between English and all 22 Indian languages, as well as direct Indic-to-Indic translation across 462 language pairs. It is optimized for high accuracy, long-context translation with both large (1.1B) and distilled (211M) model variants. It is fine-tuned using the same PEFT-LoRA methodology applied to NLLB-200. Identical LoRA hyperparameters (rank $r = 64$, $\alpha = 128$) target the

Table 3: Evaluation metrics (BLEU, METEOR, ROUGE-L, chrF, and TER) for translation directions from English to four low-resource Indic languages for the evaluation dataset.

Language Pair	BLEU	METEOR	ROUGE-L	chrF	TER
en-as	17.5352	0.4223	0.0073	57.7459	71.1716
en-mni	4.1514	0.1554	0.0113	43.8669	93.1607
en-lus	15.8280	0.4193	0.5480	51.9998	69.0074
en-bodo	19.7083	0.4549	0.1694	62.4723	64.9709

Table 4: Evaluation scores (BLEU, METEOR, ROUGE-L, chrF, TER, and Cosine Similarity) for Indic-to-English translation directions for the evaluation dataset.

Language Pair	BLUE	METEOR	ROUGE-L	chrF	TER	Cosine Similarity
As-En	0.3715	0.0127	0.0224	14.2593	116.7097	0.0388
Mni-En	8.1004	0.4798	0.4947	49.5997	100.2915	0.7974
Lus-En	12.2975	0.5778	0.6198	58.1381	78.8102	0.8888

query/key/value projections and dense layers, with DORA enhancing adaptation stability. We retain the 8-bit quantization strategy and FP16 mixed-precision training, but reduce gradient accumulation steps to 2 (effective batch size 8) due to the model's smaller footprint. The cosine learning rate schedule (3×10^{-5} peak, 500 warmup steps) and AdamW fused optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$) mirror the NLLB configuration, as does the BLEU-optimized checkpointing regime. Dynamic batching via `DataCollatorForSeq2Seq` maintains padding efficiency, while disabled caching ensures memory headroom on T4 GPUs. This consistent approach allows fair comparison between the two SOTA multilingual systems while respecting their architectural differences.

In our evaluation pipeline, we adopt a systematic approach to compute evaluation metrics for assessing the translation quality of the two models. Before evaluation, the model-generated text is preprocessed once more to enhance the reliability of metric computation for the target language. After obtaining predictions and corresponding reference labels, both sequences are decoded using the tokenizer, with special tokens skipped during decoding. To ensure compatibility with BLEU and other metrics and to correctly handle padding label tokens, marked as -100 are replaced with the tokenizer's padding token ID. A key component of our implementation is the use of the `indic_transliteration` library (Kunchukuttan, 2020b), which converts the predicted text into the appropriate target language script. This transliteration step is crucial because, in the case of the **IndicTrans**2 model, the outputs are internally generated in the Devanagari script. In contrast, the reference translations are provided in native Indic scripts. Without this conversion, evaluation metrics would be skewed due to script mismatches rather than actual translation errors. Following transliteration, the decoded sequences are post-processed by removing extraneous whitespace, and evaluation is carried out using HuggingFace's `evaluate` toolkit (Lhoest et al., 2021), which provides robust and script-aware translation metrics for Indic languages.

5 Results and Discussion

We evaluate the translation quality of the fine-tuned models using a suite of established automatic evaluation metrics, with results presented in Tables 3 and 4. These results offer key insights into the relative difficulty and success of translating between English and four underrepresented Indic languages (i.e., Assamese, Manipuri, Mizo, and Bodo) in both directions.

Table 3 reports the evaluation results for English-to-Indic translation across four low-resource languages: Assamese, Manipuri, Mizo, and Bodo. Among these, the English-to-Bodo direction achieves the highest scores across multiple metrics, BLEU (19.70), METEOR (0.4549), and chrF (62.47), indicating superior translation adequacy and fluency under the proposed approach. For final output generation, model selection was based on a comparative analysis of evaluation scores obtained

from IndicTrans2 and NLLB-200. The results show that NLLB-200 consistently outperforms IndicTrans2 for English-to-Assamese, Manipuri, and Mizo translations, whereas for the English-to-Bodo direction, IndicTrans2 demonstrates a clear advantage, yielding better translation quality.

Table 4 presents the evaluation metrics for translations from Indic languages to English. Among the language pairs, the Lus-En direction exhibits the strongest performance across nearly all metrics, BLEU (12.29), METEOR (0.5778), ROUGE-L (0.6198), chrF (58.13), and cosine similarity (0.8888), indicating high lexical and semantic alignment. In this translation direction, it was observed that the NLLB-200 model consistently outperforms IndicTrans2 for all three languages: Assamese, Manipuri, and Mizo.

6 Conclusion

This study presents a comprehensive investigation into improving machine translation quality for low-resource Indic languages through parameter-efficient fine-tuning of large multilingual models. Leveraging LoRA and DoRA techniques, we fine-tuned both the NLLB-200 and IndicTrans2 models on a curated and rigorously filtered WMT2025 dataset. Our extensive preprocessing pipeline, tailored to address the idiosyncrasies of Indic languages, proved essential in ensuring clean and semantically aligned parallel corpora. The empirical results underscore that while NLLB-200 exhibits superior performance across most language pairs and metrics, especially in English-to-Indic and Indic-to-English directions involving Assamese, Manipuri, and Mizo, IndicTrans2 offers competitive results and even outperforms NLLB-200 in the English-to-Bodo direction.

Notably, our integration of script-aware post-processing and selective transliteration was instrumental in achieving faithful metric evaluations, avoiding script mismatch penalties that would otherwise misrepresent model performance. These findings not only validate the efficacy of LoRA-based adaptation in low-resource settings but also highlight the value of task-specific linguistic preprocessing for Indic languages. Our comparative benchmarking, involving multiple metrics, reveals the nuanced translation difficulty across language pairs and emphasizes the importance of direction-aware evaluations in multilingual NMT research.

Limitations

The WMT 2025 corpora, while suitable for benchmarking, are inherently limited in scale and domain diversity for certain language pairs, particularly English–Bodo and English–Manipuri. This scarcity restricts the models' ability to generalize to informal, noisy, or domain-specific contexts.

Although the preprocessing pipeline is comprehensive, fixed thresholds in semantic filtering and transliteration heuristics may inadvertently remove valid rare sentences or alter named entities. Subtle linguistic phenomena such as dialectal variation and code-mixing remain insufficiently addressed.

Methodologically, the study is restricted to LoRA and DoRA-based fine-tuning of NLLB-200 and IndicTrans2. Although this approach ensures parameter-efficient adaptation, it does not investigate other model architectures or combined training strategies that may more effectively address unique linguistic characteristics. Similarly, the exclusive use of automatic metrics provides reproducible benchmarks but offers limited insight into true semantic quality or culturally appropriate translations.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. 2016. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*.

Raj Dabre et al. 2022. Indicbart: A pre-trained model for indic languages. In *Proceedings of LREC*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

878–891, Dublin, Ireland. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021a. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Edward J. Hu et al. 2021b. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Melvin Johnson, Mike Schuster, Quoc V Le, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. In *Transactions of the Association for Computational Linguistics*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english–nyishi pair. *Sādhanā*, 48(4):237.

Philipp Koehn. 2005a. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86.

Philipp Koehn. 2005b. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anoop Kunchukuttan. 2020a. The indic nlp library. Accessed: 2025-07-30.

Anoop Kunchukuttan. 2020b. The indic nlp library. Accessed: 2025-07-30.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2014. Sata anuvadak: Tackling multiway translation for indian languages. In *Proceedings of WAT*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Quentin Lhoest, Benjamin Minixhofer, Siddhartha Bandyopadhyay, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu et al. 2020. Multilingual denoising pre-training for neural machine translation. In *Transactions of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *International Conference on Learning Representations (ICLR)*.

Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. Findings of WMT 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT / EMNLP 2025)*.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Sarah Lyngdoh, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Gowtham Ramesh, Vishrav Chaudhary, Divyanshu Kakwani, Sai Praneeth Golla, Abhishek Philip, et al. 2023. Indictrans2: Towards high-quality and efficient multilingual translation for indic languages. *arXiv preprint arXiv:2304.09105*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Celebi, and et al. 2022a. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

NLLB Team et al. 2022b. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In *NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, and Jamie Brew. 2020. Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu-Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *CoRR*, abs/2312.12148.

Linting Xue et al. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*.

Lianmin Zhao, Shrimai Prabhumoye, Chen Shao, Yihong He, Xiaodong Ma, et al. 2023. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2306.11695*.

An Attention-Based Neural Translation System for English to Bodo

Subhash Kumar Wary[1], **Birhang Borgoyary**[2], **Akher Uddin Ahmed**[3], **Mohanji Prasad Sah**[4], **Apurbalal Senapati**[5]
[1,2,3,4,5]Central Institute of Technology Kokrajhar
BTR Assam, India, Pin - 783370
subhashkumarwary@gmail.com, bborgoyary021@gmail.com, akheruddinahmedcse@gmail.com, mohanjiprasadsah80@gmail.com, a.senapati@cit.ac.in

Abstract

Bodo is a resource scarce, the indigenous language belongs to the Tibeto-Burman family. It is mainly spoken in the north-east region of India. It has both linguistic and cultural significance in the region. Only a limited number of resources and tools are available in this language. This paper presents a study of neural machine translation for the English-Bodo language pair. The system is developed on a relatively small parallel corpus provided by the Low-Resource Indic Language Translation as a part of WMT-2025 [1]. The system is evaluated by the WMT-2025 organizers with the evaluation matrices like BLUE, METEOR, ROUGE-L, chrF and TER. The result is not promising but it will help for the further improvement. The result is not encouraging, but it provides a foundation for further improvement.

1 Introduction

The Bodo language, which belongs to the Sino-Tibetan language family, is one of the widely spoken languages in Assam and several other parts of the North-Eastern states of India. It is used predominantly in the Bodoland Territorial Region (BTR), which includes the districts of Kokrajhar, Chirang, Baksa, and Udalguri, as well as in other districts such as Kamrup, Sonitpur, Lakhimpur, Nagaon, Morigaon, and Karbi Anglong. Bodo is one of the 22 languages listed in the Eighth Schedule of the Indian Constitution and is officially recognized by the Government of India (Census, 2011). According to the 2011 Census, it is spoken by more than a million people, primarily members of the Bodo community (Koyel Ghosh, 2023). The number of Bodo speakers is shown in the Figure 1. The Bodo language has rich linguistic features and uses the Devanagari script for writing, similar to Hindi. Having the tonal feature. Hence, effective techniques are not developed to capture all these features (Mwnthai Narzary, 2022).

Machine translation is a core application in the field of Natural Language Processing (NLP). With advancements in computational power, the focus has shifted from rule-based methods to machine learning and deep learning approaches. However, implementing deep learning techniques requires a large volume of data (Narzary Sanjib, 2019). In this paper, we have developed an English-Bodo machine translation system using a transformer-based neural machine translation approach. Preprocessing steps such as tokenization, subword extraction, and normalization are required before feeding the data into the actual Transformer model (Guillaume et al., 2017).

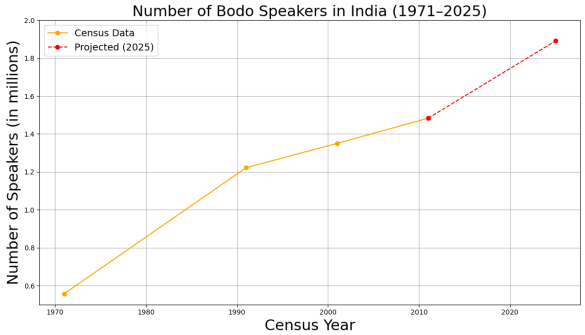


Figure 1: Number of Bodo speakers in India from 1971 to 2025 (projected).

2 Related Work

As mentioned above, Bodo is a low-resource language (Narzary et al., 2021), and development a machine translation system for it faced significant challenges due to the limited availability of parallel corpora and digital resources (Kalita et al., 2023). Most research efforts have focused on creating and expanding parallel corpora, as well as adapting machine translation techniques to function effectively

[1]https://www2.statmt.org/wmt25/indic-mt-task.html

Proceedings of the Tenth Conference on Machine Translation, pages 1210–1214
November 8-9, 2025

with scarce data. Although work in Bodo machine translation remains limited, there have been some notable efforts, particularly in building English-Bodo translation systems. A brief overview of these works is provided below.

Bahdanau et al. (Dzmitry Bahdanau, 2015) observed that traditional neural machine translation and statistical machine translation models have an alignment problem that affects the performance. The common encoder-decoder architecture, which uses an encoder to compress a source sentence into a single, fixed-length vector, faced a critical bottleneck: this fixed-length vector was often insufficient to capture all the information from a long sentence. To solve this, they proposed a new model that allows the decoder to automatically search for and focus on the most relevant parts of the source sentence when predicting each word of the translation. This mechanism, known as attention, significantly improved translation quality by overcoming the limitations of a single context vector. Verma et al. (Verma and Bhattacharyya, 2017) conducted a literature survey on Neural Machine Translation (NMT), highlighting the advantages of the NMT architecture. Vaswani et al. (Vaswani et al., 2017) proposed the attention-based Transformer model, which gained significant popularity in machine translation due to its novel architecture and promising performance. Parvez et al. (Parvez et al., 2023) attempted neural machine translation for the pair of low-resource languages of English and Bodo. They utilized a relatively small English-Bodo parallel corpus and implemented their system using the OpenNMT-py framework. Their model achieved a highest BLEU score of 11.01. Islam et al. (Islam and Purkayastha, 2019) worked on Bodo-to-English machine translation using a phrase-based statistical machine translation (PB-SMT) approach. They applied this technique to Bodo-English parallel corpora in both the general and news domains, reporting a highest BLEU score of 30.13. Narzary et al. (Narzary Sanjib, 2019) developed an attention-based English-Bodo neural machine translation system using data from the tourism domain. Their baseline model achieved a BLEU score of 11.8. By incorporating an attention mechanism, they improved the model's performance, reaching a BLEU score of 16.71. Talukdar et al. (Talukdar et al., 2023) focused on Assamese-Bodo neural machine translation and investigated the impact of data quality and quantity on translation performance. They iteratively augmented the dataset and evaluated the outcomes at each stage. The experiments were conducted using the OpenNMT-py framework. Gaikwad et al. (Gaikwad et al., 2024) suggested that the use of a high-resource language as a pivot can improve translation into related low-resource languages. They conducted experiments on machine translation of the English to Indian language - specifically translating English into Konkani, Manipuri, Sanskrit, and Bodo - employing Hindi, Marathi, and Bengali as pivot languages.

3 System Description

We have implemented the Attention-Based Neural Machine Translation (NMT) system. It is a deep learning model specifically designed for sequence-to-sequence tasks, such as translating text from one language to another. In traditional sequence-to-sequence models, the entire input sentence is encoded into a single, fixed-length vector that captures its relevant contexts. That single vector cannot effectively capture all the rich context of long or complex sentences. On the other hand, attention-based architecture allows the model dynamically to focus on relevant parts of the input sentence while generating each word in the output. The system typically consists of an encoder-decoder architecture along with an attention mechanism. The attention mechanism dynamically modifies the context vector for each output word. This allows the decoder to "attend" to different parts of the input sentence at each step. The basic encoder-decoder architecture, along with the attention mechanism, is depicted in Figure 2. The figure is configured for the English-Bodo translation, which is influenced by the Tato et al. (Tato and Nkambou, 2022) diagram.

The attention mechanism used in the decoder to decide which parts of the input sequence to focus on while generating an output. Calculating a context vector by taking a weighted sum of the encoder's hidden states. The weights for this sum are dynamically adjusted, giving more importance to the input words that are most relevant to the current output being generated. Based on this mechanism, the model can capture long-range dependencies and produce higher-quality translations. This is particularly effective for long or complex sentences or when translating between languages with different word orders.

In the attentional model (Dzmitry Bahdanau,

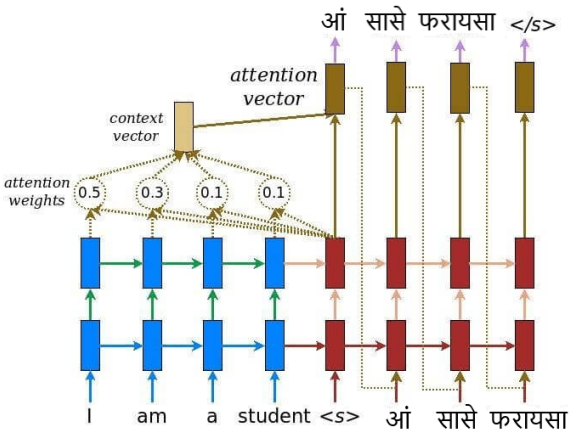


Figure 2: Example of attention mechanism in the translation from English to Bodo

2015), all hidden states h_i of the encoder are utilized to compute the context vector c_t. This model generates a variable-length alignment vector a_t, whose size corresponds to the number of time steps on the source side. The alignment vector is obtained by comparing the current hidden state h_i with each encoder hidden state $\bar{h_s}$. Where:

$$\alpha_{ts} = \frac{exp(score(h_t, \bar{h_s}))}{\sum_{s'} exp(score(h_t, \bar{h_{s'}}))} \quad (1)$$

$$c_t = \sum \alpha_{ts} \bar{h_s} \quad (2)$$

$$a_t = f(c_t.h_t) = tanh(W_c[c_t; h_t]) \quad (3)$$

$$a_t = f(c_t.h_t) = tanh(W_c[c_t; h_t]) \quad (4)$$

The $score(h_t, \bar{h_s})$ is calculated as

$$\begin{cases} {h_t}^T W \bar{h_s} & \text{[Luong]} \\ {v_a}^T tanh(W_1 h_t + W_2 h_s) & \text{[Bahdanau]} \end{cases}$$

Here, both the multiplicative and additive (Luong et al., 2015) (Dzmitry Bahdanau, 2015) attention mechanisms have normalized variants, known respectively as Scaled Luong and Normed Bahdanau (Raffel et al., 2017). The main idea behind the attention vector is to determine how much emphasis should be focused on each source word at a given time step. A higher value in the attention weight at indicates that the corresponding source word has a greater influence on predicting the next word in the output sentence. Particularly, the model the model performs translation based on the conditional probability $p(y|x)$, which represents the likelihood of translating a source sentence $x_1; ....; x_n$, into a target sentence $y_1; ....; y_m$. This is achieved using an encoder-decoder framework.

4 Dataset used

The data set we used in this work for the translation task is provided in the Very Limited Training Data setting of the WMT25 Indic Multilingual Machine Translation Shared Task [2], a snapshot of the data is shown in Figure 3 and Figure 4 respectively. Bodo is a low-resource language, and the availability of high-quality parallel corpora is significantly constrained. The official data set provided by the WMT25 organizers consists of a small number of English-to-Bodo sentence pairs curated for the purpose of benchmarking machine translation systems under low resource conditions. In the dataset, contains training, development, and test splits, with the training set including only a few thousand sentence pairs. These sentences span general purpose domains such as basic conversational language. The development set was used for validation and tuning, while the test set was reserved for blind evaluation by the task organizers.

We have also taken a parallel data set focused on tourism [3] augmented with WMT25 data for training purposes shown in Table 1. This data set contains English and Bodo corpus, which we trained and tested in a different model for the contrastive output element. The data set is divided into six subsets to facilitate training, validation, and testing for both languages involved in translation. Specifically, they have provided *train_brx, val_brx*, and *test_brx* contain Bodo sentences for training, validation, and testing respectively, while *train_eng, val_eng*, and *test_eng* contain the corresponding English sentences. Each Bodo sentence in a given split aligns with its English counterpart, enabling parallel corpus training for machine translation models.

Sl. No.	Corpus name	# Files	# Sentences
1	WMT25 (en-bodo)	1	15,216
2	Tourism (en-bodo)	6	33,258

Table 1: English-Bodo training data set

[2]https://www2.statmt.org/wmt25/indic-mt-task.html
[3]https://get.alayaran.com/parallel-data/

English
The Indian independence movement was a significant period in Indian history, marked by a fervent desire for freedom from British rule.
Mahatma Gandhi emerged as the leader of the Indian independence movement, advocating for nonviolent resistance against British colonial rule.
The Salt March of 1930, led by Mahatma Gandhi, was a pivotal event in the Indian independence movement, protesting the salt tax imposed by the British.
The Quit India Movement, launched in 1942, was a major civil disobedience movement aimed at demanding an end to British rule in India.
Jawaharlal Nehru, the first Prime Minister of independent India, played a crucial role in the Indian independence movement and the nation's subsequent develo
The Partition of India in 1947 led to the creation of two separate nations, India and Pakistan, resulting in one of the largest mass migrations in human history.
The Indian National Congress, founded in 1885, played a central role in the Indian independence movement, advocating for self-rule and independence from Br
Bhagat Singh, a revolutionary freedom fighter, became an iconic figure in the Indian independence movement through his acts of protest against British rule.
The Jallianwala Bagh massacre of 1919 was a tragic event during the Indian independence movement, where British troops indiscriminately killed hundreds of
The Indian National Army (INA), led by Subhas Chandra Bose, fought alongside the Japanese during World War II, seeking to liberate India from British rule.

Figure 3: WMT25 Data set - English (eng)

Bodo
भारतनि उदांस्त्रि सोमावसारनाया भारतनि जारिमिनाव मोनसे मख'जाथाव सममोन, जाय ब्रिटिश खुंथायनिफ्राय उदांस्त्रिनि गोसो गुदुं लुबैनायजों सिनायथि होजादोंमोन।
महात्मा गान्धीआ भारतारि उदांस्त्रि सोमावसारनायनि दैदेनगिरि महरै नुजादोंमोन, जाय ब्रिटिश फामुआरि खुंथायनि बेरेखायै अहिंसा होबथानायनि वकालत खालामदोंमोन।
महात्मा गान्धीनि दैदेन्नायाव 1930 मायथाइनि संख्रि हान्जा सुरनाया भारतनि उदांस्त्रि सोमावसारनायाव मोनसे गोनांथार हाबाफारिमोन, जाय इंराजफोरजों होजानाय संख्रि खाजोनाखौ हेंथा होदोंमोन।
1942 मायथाइयाव जागायजेन्नाय भारत नागार सोमावसारनाया भारतआव ब्रिटिश खुंथायखौ फोजोबनो दाबि खालामनायनि थांखिजों मोनसे गाहाय नोगोरारि मानियै सोमावसारनायमोन।
उदां भारतनि गिबि गाहाय मन्थ्रि जवाहरलाल नेहरुआ भारतनि उदांस्त्रि सोमावसारनाय आरो हादोरनि उननि जौगानायाव गोनांथार बिफाव लादोंमोन।
1947 मायथाइयाव भारतनि बोखावनाया मोन्नै गुबुन गुबुन हादोरफोर, भारत आरो पाकिस्ताननि जोनोम होदोंमोन, जायनि फिथाय महरै सुबुं जारिमिनाव बयनिख्रुइबो देरसिन सुबुं हाबफैनायफोरनि गेजेराव मोनसे।
1885 मायथाइयाव गायसनजानाय भारतारि माहारियारि कंग्रेसा भारतनि उदांस्त्रि सोमावसारनायाव मोनसे मिरुआरि बिफाव लादोंमोन, जाय ब्रिटिश फामुवारिबादनिफ्राय गावारि-सासन आरो उदांस्त्रिनि हेफाजाब होदोंमोन।
सासे बिग्रायारि उदांस्त्रि सोमावसारगिरि भगत सिंहआ ब्रिटिश खुंथायनि बेरेखायै गावनि हेंथा दिन्थिफुंनायनि गेजेरजों भारतनि उदांस्त्रि सोमावसारनायाव सासे मुंदांखा सुबुं जादोंमोन।
1919 मायथाइनि जालियानवाला बाग बुथारग्लुंनाया भारतनि उदांस्त्रि सोमावसारनायनि समाव मोनसे दुखुनांथाव जाथायमोन, जेराव ब्रिटिश रौनियाफोरा अमृतसरआव जैरै-मैरै गोबां हाथियार गैयै नोगोरारिफोरखौ बुथारदोंमोन।
सुभाष चन्द्र ब'सनि दैदेन्नायाव भारतारि हायुंआरि रौनिया हानजाया (आई.एन.ए.) नैथि मुलुग दावहानि समाव जापानिफोरजों लोगोसे दावहा नांदोंमोन, भारतखौ ब्रिटिश खुंथायनिफ्राय उदां खालामनो नाजादोंमोन।

Figure 4: WMT25 Data set - Bodo (brx)

5 Implementation

All model training and experimentation were conducted using Google Colab, a cloud-based development platform. We utilized the NVIDIA A100 GPU available through Google Colab for training our NMT model. The A100, based on NVIDIA's Ampere architecture, offers high memory bandwidth and massive parallel processing capabilities, making it a well suited for deep learning tasks. Its support for mixed precision training and large batch processing significantly accelerated model training and testing. The GPU enabled efficient handling of our encoder to decoder architecture with attention which allowed us to train and test on the English to Bodo dataset with reduced computation time and improved performance.

The model is developed and trained using PyTorch within a Google Colab environment. The data set is cleaned by removing null values and then shuffled to eliminate order bias. We load a test set provided as plain text files, ensuring that sentence alignment is preserved across the splits of training and validation sets. For preprocessing, we utilize a custom LanguageProcessor class that tokenizes the data and constructs a vocabulary with special tokens <PAD>, <SOS>, <EOS>, and <UNK>. It maps words to indices and vice versa and computes the maximum sentence length for padding.

We configured a sequence-to-sequence encoder-decoder model with attention using carefully selected hyperparameters to ensure efficient training and optimal performance. The embedding dimension for both the encoder and decoder was set to 256, while the hidden state dimensions were set to 512 units. The dropout regularization was applied with a rate of 0.5 in both the encoder and decoder. Training was conducted using a batch size of 64 over 15 epochs.

6 Result

The system is evaluated by the WMT25 organiser, which provided a test dataset of 1,000 sentences. A total of nine runs were assessed for the task. Five evaluation metrics were used: BLEU, METEOR, ROUGE-L, chrF, and TER[4]. The result of our system is shown on Table 2. For comparison, the highest and lowest scores for the track are presented in Table 3 and Table 4, respectively.

Sl. No.	Metric	Score
1	BLEU	0.3106045292
2	METEOR	0.01875594452
3	ROUGE-L	0.002595238095
4	chrF	7.235394682
5	TER	808.9101286

Table 2: Result of the system (en-bodo)

Sl. No.	Metric	Score
1	BLEU	24.44868688
2	METEOR	0.5126346512
3	ROUGE-L	0.1684904762
4	chrF	67.70727358
5	TER	51.84296487

Table 3: Highest score of the track (en-bodo)

Sl. No.	Metric	Score
1	BLEU	0.2047914219
2	METEOR	0.006037416908
3	ROUGE-L	0.02716098904
4	chrF	0.8138721309
5	TER	131.9585726

Table 4: Lowest score of the track (en-bodo)

7 Conclusion

It is observed that your result (Table 2) is not much surprising. While the score is higher than the lowest score (Table 4), it is still lower compared to the highest (Table 3). To investigate its weaknesses, a granular-level error analysis is needed. At a glance, we found that the system performs poorly on complex sentences compared to simple ones. The data

[4]https://www2.statmt.org/wmt25/mteval-subtask.html

provided by the track, along with the various systems presented here, will be valuable for future research.

Limitations

The training dataset is insufficient for developing a sophisticated machine translation system.

Ethics Statement

Not Applicable

Acknowledgements

We sincerely thank Mr. Sanjib Narzary for his valuable guidance in implementing the system.

References

Census. 2011. Abstract of speakers' strength of languages and mother tongues – census 2011. *Office of the Registrar General Census Commissioner, India. 2018*, New Delhi: Ministry of Home Affairs, Government of India.

Yoshua Bengio Dzmitry Bahdanau, Kyunghyun Cho. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*.

Pranav Gaikwad, Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. How effective is multi-source pivoting for translation of low resource indian languages? *ArXiv*, abs/2406.13332.

Klein Guillaume, Kim Yoon, Yuntian Deng, Senellart Jean, and Rush Alexander. 2017. OpenNMT: Open-source toolkit for neural machine translation. pages 67–72.

Saiful Islam and Bipul Syam Purkayastha. 2019. Bodo to english machine translation through transliteration. *International Journal of Innovative Technology and Exploring Engineering*.

S. Kalita, P. Boruah, Kishore Kashyap, and Shikhar Kr Sarma. 2023. Nmt for a low resource language bodo: Preprocessing and resource modelling. *2023 4th International Conference on Computing and Communication Systems (I3CS)*, pages 1–5.

Mwnthai Narzary Maharaj Brahma Koyel Ghosh, Apurbalal Senapati. 2023. Hate speech detection in low-resource bodo and assamese texts with ml-dl and bert models. *Scalable Computing: Practice and Experience*, 24(4).

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Sanjib Narzary Apurbalal Senapati Pranav Kumar Singh Mwnthai Narzary, Maharaj Brahma. 2022. A computational approach for the tonal identification in bodo language. *Bhattacharjee, R., Neog, D.R., Mopuri, K.R., Vipparthi, S.K. (eds) Artificial Intelligence and Data Science Based RD Interventions. NERC 2022*.

Mwnthai Narzary, Gwmsrang Muchahary, Maharaj Brahma, Sanjib Narzary, P. Singh, and Apurbalal Senapati. 2021. Bodo resources for nlp - an overview of existing primary resources for bodo. *Proceedings of Intelligent Computing and Technologies Conference*.

Singha Bobita Brahma Rangjali Dibragede Bonali Barman Sunita Nandi Sukumar Som Bidisha Narzary Sanjib, Brahma Maharaj. 2019. Attention based english-bodo neural machine translation system for tourism domain. pages 335–343.

Boruah Parvez, Talukdar Kuwali, Ahmed Mazida, and Kashyap Kishore. 2023. Neural machine translation for a low resource language pair: English-bodo.

Colin Raffel, Minh-Thang Luong, Peter J. Liu, Ron J. Weiss, and Douglas Eck. 2017. Online and linear-time attention by enforcing monotonic alignments. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2837–2846.

Kuwali Talukdar, Shikhar Kumar Sarma, Farha Naznin, and Kishore Kashyap. 2023. Influence of data quality and quantity on assamese-bodo neural machine translation. *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5.

Ange Tato and Roger Nkambou. 2022. Infusing expert knowledge into a deep neural network using attention mechanism for personalized learning environments. *Front. Artif. Intell*, 5:921476.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Ajay Anand Verma and Pushpak Bhattacharyya. 2017. Literature survey: Neural machine translation. *CFILT, Indian Institute of Technology Bombay, India*.

Tackling Low-Resource NMT with Instruction-Tuned LLaMA: A Study on Kokborok and Bodo

Deepak Kumar, Kshetrimayum Boynao Singh, Asif Ekbal
Indian Institute of Technology, Patna
{deepakkumar1538, boynfrancis, asif.ekbal}@gmail.com

Abstract

This paper presents a new neural machine translation (NMT) system aimed at low-resource language pairs: English to Kokborok and English to Bodo. The framework leverages the LLaMA3-8B-Instruct model along with LoRA-based parameter-efficient fine-tuning. For translating into Kokborok, the model undergoes an initial continued pre-training phase on a dataset containing 75,000 Kokborok and 25,000 English monolingual sentences, followed by instruction-tuning. This tuning uses a reformulated version of WMT25 dataset, adapted to the Alpaca format to support instructional goals. In the Bodo translation, the model is pre-trained on a more extensive dataset of 350,000 Bodo and 125,000 English sentences, using a similar instruction-tuning approach. LoRA adapters are used to modify the large LLaMA3 model for these low-resource settings. Testing with the WMT25 test dataset reveals modest translation results, highlighting the difficulties in translating for low-resource languages. Translating English to Bodo, the model achieved a BLEU score of 4.38, a TER of 92.5, and a chrF score of 35.4. For English to Kokborok, it yielded scores of 5.59 in chrF, 105.4 in TER, and 0.17 in BLEU. These results underscore the intricacies of the task and highlight the critical need for further data collection, domain-specific adaptations, and improvements in model design to better support underrepresented languages.

1 Introduction

Despite significant advancements in neural machine translation (NMT) and the instruction-tuning of large language models (LLMs), the predominant focus of research and datasets remains on English and high-resource languages. Instruction-tuning, a potent method to align LLMs with human preferences, has demonstrated efficacy primarily in contexts where English data are plentiful. Conversely, low-resource languages frequently experience a dearth of robust foundational models due to the scarcity of both monolingual and parallel corpora. In mitigating this limitation, recent studies investigate cross-lingual instruction-tuning by integrating translation-following demonstrations, which enable English-centric LLMs to generalize to novel languages. Currently, evidence suggests that zero-shot cross-lingual transfer is achievable when instruction-tuning is meticulously calibrated, even in scenarios where only English instructions are employed, although factuality and fluency may be compromised in the target language.

More comprehensively, multistage architectures, such as LinguaLIFT, further exemplify how code-switched translation data and task alignment can enhance reasoning in low-resource environments without dependence on extensive multilingual corpora. In light of these insights, this paper introduces an innovative NMT framework targeting pairs of English to Kokborok and English to Bodo language, utilizing the LLaMA3-8B-Instruct model, fine-tuned with LoRA-based adapters for efficient parameter adaptation. The model is initially pre-trained on monolingual corpora of 75,000 Kokborok and 25,000 English sentences for Kokborok and 350,000 Bodo and 125,000 English sentences for Bodo, subsequently undergoing instruction-tuning on WMT25-supplied English–target language parallel data reformatted into an Alpaca style structure. This process aligns translation objectives with instruction follow-up behaviors within the LLM.

Evaluation in WMT25 test sets reveals modest yet meaningful translation performance: for English to Bodo, BLEU = 4.37, TER = 92.5, and chrF = 35.4; for English to Kokborok, chrF = 5.59, TER = 105.4, and BLEU = 0.17. These findings highlight the considerable challenges inherent in low-resource machine translation and emphasize the crucial need for persistent data collection, domain adaptation, and alignment strategies.

Proceedings of the Tenth Conference on Machine Translation, pages 1215–1221

Our key contributions are:

- We present a **LoRA-adapted fine-tuning methodology** to enhance English centric, instruction-following models (LLaMA) for translation tasks in low-resource languages such as Kokborok and Bodo.
- We propose a **sequential pipeline** combining monolingual *pre-training* followed by *instruction-tuning*, aimed at strengthening language representations for underrepresented Indian languages.
- We conduct an **empirical evaluation** that highlights both the limitations and capabilities of instruction-tuned LLaMA models in extremely low-resource translation scenarios.
- We issue a **call for future work** aiming at the expansion of high-quality instruction-tuning datasets, domain adaptation techniques, and architecture-level innovations to support the broader inclusion of under-resourced languages in LLM pipelines.

2 Linguistic Background

In this study, we focus on two low-resource Indian languages, Bodo and Kokborok, both belonging to the Tibeto-Burman branch of the Sino-Tibetan language family. These languages are spoken primarily in the northeastern region of India and share several linguistic characteristics such as tonality, agglutinative morphology, and subject object verb (SOV) word order. Despite their socio-cultural and political significance, both languages remain significantly underrepresented in NLP research and lack sufficient digital resources for computational modeling.

2.1 Bodo Language

Bodo[1], also spelled Boro, is a low-resource language spoken predominantly in the Bodoland Territorial Region of Assam, India. According to the 2011 Indian Census, it has more than 1.4 million speakers and is officially recognized as one of the 22 scheduled languages of India. Bodo serves as a medium of instruction in schools throughout the region and has official status in the local administration. The language has historically used Latin and Bengali scripts, but now primarily uses Devanagari. Linguistically, Bodo follows an order of words SOV and exhibits features such as tonality, agglutination, and rich verbal morphology, including case marking and verb inflections. Despite its cultural significance, Bodo lacks substantial computational resources, making it an important target for low-resource NLP, particularly in areas such as neural machine translation (MT), language modeling, and morphological analysis. Its inclusion in multilingual NLP initiatives contributes both to linguistic inclusivity and to technological equity.

2.2 Kokborok Language

Kokborok[2], also known as Tripuri, is another Tibeto-Burman language of the Sino-Tibetan family, spoken mainly in the state of Tripura and parts of Bangladesh. It is the mother tongue of the Tripuri people and one of the most spoken dialects among the various linguistic varieties of the Tripuri. According to the 2011 Census, it has more than 1 million speakers. Kokborok lacks a standardized orthography; however, the Roman script is increasingly used in digital communication and education, along with some continued use of Bengali and Devanagari scripts based on socio-political and institutional preferences. Like Bodo, Kokborok follows an SOV syntactic structure, is tonal, and agglutinative, with a complex system of verbal inflection and case marking. Despite its status as one of the official languages of Tripura, Kokborok is largely lacking digital resources, with very limited availability of annotated corpora, linguistic tools, or parallel datasets. This presents significant challenges for NLP system development, but also offers an opportunity to explore transfer learning, few-shot adaptation, and cross-lingual methods in the context of extremely low-resource settings.

3 Related Work

3.1 Low-Resource MT for Indic Languages

The domain of low-resource neural machine translation (NMT) has been extensively studied with respect to underrepresented Indic languages through the use of a variety of methodologies (Kakwani et al., 2020). In this context, English to Bodo translation systems constructed using approximately 92K parallel corpus sentences with Transformer-based architectures, thereby achieving BLEU scores reaching 11.01.(Boruah et al.,

[1] https://en.wikipedia.org/wiki/Boro_language_(India)

[2] https://en.wikipedia.org/wiki/Kokborok

2023). Similar efforts for languages such as Manipuri (Singh et al., 2023b, 2024), Assamese, Mizo, and Khasi have employed techniques such as transfer learning (Singh et al., 2023a) and tokenization of sub-words to address challenges associated with data sparsity.

3.2 Pre-training and Instruction-Tuning of LLMs

The adaptation of large language models (LLMs) for resource-constrained languages, achieved (Gao et al., 2024) via continued monolingual pre-training (Sennrich et al., 2016) followed by instruction-tuning, has shown favorable results. Specifically, presents an Estonian instruction follow-up LLM derived from LLaMA-2, showcasing that the integration of constrained monolingual pre-training with cross-lingual instruction-tuning considerably improves performance on Estonian tasks. Furthermore, they have introduced Alpaca-est, which represents the first general-task instruction dataset for the Estonian language.

3.3 Cross-Lingual Instruction-Tuning

Cross-lingual instruction-tuning where instructions and translation tasks appear in both English and the target language has been proposed as a cost efficient adaptation method for English centric LLMs (Gao et al., 2024). Such tuning is often implemented in Alpaca style formats to align translation and general task objectives. The inclusion of translation-style instructions during tuning has been empirically shown to boost performance in target languages, especially when paired with monolingual pre-training.

3.4 Parameter Efficient Fine-Tuning (PEFT)

Parameter efficient fine-tuning techniques like LoRA have become increasingly popular for adapting massive LLMs (Hu et al., 2021) to new tasks and languages under compute constraints. PEFT methods enable efficient adaptation without needing to update all model weights, making them especially appealing for low-resource settings.

3.5 Data Augmentation and Synthetic Parallel Corpora

Data augmentation methods especially back-translation remain vital for enhancing low-resource MT performance. These methods generate synthetic parallel data from monolingual corpora (Raja and Vats, 2025), though the quality depends heavily on the strength of the reverse translation model. Controlled generation techniques (e.g., using target sentence length tokens) have also been explored to manage translation output fidelity.

3.6 Evaluation Metrics and Limitations in Low-Resource MT

Evaluation for low-resource MT frequently relies on automatic metrics such as BLEU (Papineni et al., 2002), chrF (Popović, 2017), and TER (Snover et al., 2006). While COMET (Rei et al., 2020), a neural based metric, often correlates better with human judgments, its availability remains limited for languages like Bodo and Kokborok. Human evaluation is crucial to assess fluency and adequacy, particularly for morphologically complex languages.

3.7 Gap and Our Contribution

In the domain of low-resource Indic machine translation (MT), despite developments in parameter efficient fine-tuning (PEFT) techniques, instructional tuning frameworks, and data augmentation methodologies, the translation of Kokborok and Bodo employing instruction following large language models (LLMs) such as LLaMA3 with LoRA adapters remains unexplored. This investigation integrates monolingual pre-training for Bodo and Kokborok with Alpaca style instruction-tuning and LoRA fine-tuning. The study evaluates the performance of the constructed models using the WMT25 (Pakray et al., 2025, 2024; Pal et al., 2023; Kakum et al., 2023) English to Bodo and English to Kokborok datasets and notes limited BLEU (≈ 0.17 for Kokborok), chrF, and TER scores. These results highlight the considerable challenges that persist and stress the necessity for ongoing research specific to these languages.

4 Dataset Preparation

For this study, we collected a substantial amount of monolingual data in Bodo, Kokborok, and English. The dataset comprises approximately 350,000 sentences in Bodo, 75,000 sentences in Kokborok, and 450,000 sentences in English. All data was legally compliant for research use.

For pre-training, we constructed two datasets, one for English to Bodo and another for English to Kokborok, by mixing monolingual data in the following ratios:

- **English to Bodo:** 75% Bodo, 25% English

- **English to Kokborok:** 70% Kokborok, 30% English

These specific ratios were chosen to ensure that the base model becomes more familiar with Bodo or Kokborok while preserving sufficient English fluency to prevent catastrophic forgetting.

We also prepared an instruction-tuning dataset in Alpaca format. We collected WMT25 [3] parallel training data for English to Bodo and English to Kokborok and converted it into Alpaca style instruction response pairs. To enhance the robustness of the model, we maintained a variety of instruction types so that the model learned to follow multiple instruction formats. Finally, we trained two LoRA-adapted LLaMA models using these instruction-tuned datasets for English→Bodo and English→Kokborok translation.

Stage	Language Pair	Sentences
Pre-training	English (eng)	450,000
	Bodo (bodo)	350,000
	Kokborok (trp)	75,000
Fine-tuning	English to Bodo	15,215
	English to Kokborok	2,269
Testing	English to Bodo	1,000
	English to Kokborok	1,000

Table 1: Dataset statistics used for pre-training, fine-tuning, and testing. All monolingual datasets are open-source and licensed under CC BY 4.0. Fine-tuning and testing datasets were converted into instruction response format following Alpaca style instruction-tuning.

5 Methodology

5.1 Monolingual pre-training

We first continued pre-training the LLaMA3-8B-Instruct model on a mixed monolingual corpus for each target language. For English to Kokborok, we use a mix of 75,000 Kokborok sentences 75% and 25,000 English sentences 25% and for English to Bodo, 350,000 Bodo sentences and 125,000 English sentences 75%–25%. These data sets are concatenated and shuffled to increase exposure to the target language while retaining English fluency and minimizing catastrophic forgetting.

Pre-training is conducted using the standard autoregressive language modeling objective, with next token prediction as the training objective. This stage helps the model internalize the vocabulary, grammar, and patterns of the low-resource language before instruction-tuning. As demonstrated in earlier studies (Kuulmets et al., 2024), this step significantly boosts translation performance in low-resource settings.

Following monolingual adaptation, we perform instruction fine-tuning using the WMT25 parallel corpora, consisting of 15,215 sentence pairs for English to Bodo and 2,269 pairs for English to Kokborok. Each sentence pair is reformatted into an Alpaca style instruction response format:

> **Instruction:** Translate the following sentence into [Target Language].
> **Input:** [English Sentence]
> **Output:** [Reference Translation]

This prompt format aligns with the instruction following capabilities of LLaMA-style models and encourages better alignment between English inputs and target-language outputs. No auxiliary tasks or instructions (e.g., summarization or question answering) are included; the dataset is entirely translation-focused.

5.2 LoRA Fine-Tuning

To efficiently adapt the large model to these instruction-tuned datasets, we use Low-Rank Adaptation LoRA (Hu et al., 2021). LoRA adapters are injected into each transformer layer of the model, enabling parameter-efficient fine-tuning. Only low-rank update matrices are trained, while the original model weights remain frozen.

This significantly reduces computational cost and memory usage while maintaining high translation quality. The training objective is the standard cross-entropy loss between predicted tokens and the reference translation, without any post-processing or auxiliary losses.

5.3 Experimental Infrastructure

All pre-training and fine-tuning experiments were conducted on NVIDIA A100 80 GB PCIe GPUs, deployed in a dual-GPU configuration. Each GPU offers up to 80 GB of HBM2e memory with a maximum memory bandwidth of approximately 1.9 - 2.0 TB, enabling rapid data movement, essential for training large-scale models.

This powerful GPU setup provides the computational and memory resources necessary to efficiently pretrain and finetune large language models for low-resource machine translation scenarios.

[3] https://www2.statmt.org/wmt25/indic-mt-task.html

Direction	Setting	BLEU (↑)	METEOR (↑)	ROUGE-L (↑)	chrF (↑)	TER (↓)
en-bodo	Zero-Shot	0.2	0.0060	0.0117	0.7	113.5
	Few-Shot	0.4	0.0098	**0.0201**	0.9	130.5
	Proposed	**4.38**	**0.1318**	0.0090	**35.50**	**92.56**
en-trp	Zero-Shot	0.0	0.0011	0.0023	10.2	155.1
	Few-Shot	0.1	**0.0090**	0.0087	**18.0**	158.9
	Proposed	**0.18**	0.0063	**0.0153**	5.60	**105.49**

Table 2: Comparison of Zero-Shot, Few-Shot, and Proposed Architecture Results. Bold indicates best performance per metric.

The English → Kokborok and English → Bodo experiments were trained for 10 epochs with small batch sizes, a fine-tuning learning rate of 1.5e-5, and regularization through weight decay and warmup. A cosine restart scheduler and gradient clipping ensured stable optimization, while bf16 precision improved speed and memory efficiency, making the setup well-suited for low-resource translation tasks.

6 Results and Analysis

We compare our proposed architecture against zero-shot and few-shot prompting baselines, where the few-shot setup uses the limited bilingual examples from the WMT 2025 shared task as in-context demonstrations. Table 2 reports results across BLEU, METEOR, ROUGE-L, chrF, and TER for English→Bodo (en-bodo) and English→Kokborok (en-trp).

For en-bodo, the proposed model achieves a BLEU score of 4.38, over 10× higher than both zero-shot (0.2) and few-shot (0.4). METEOR similarly improves to 0.1318, compared to 0.0060 and 0.0098. While ROUGE-L (0.0090) trails the few-shot baseline (0.0201), chrF rises sharply to 35.50, far above prompting (0.7/0.9), and TER drops to 92.56, indicating far fewer edits than either baseline. These gains show that our architecture captures structure and meaning beyond what prompting alone enables.

For en-trp, improvements are more modest but still consistent. The model reaches a BLEU of 0.18 (vs. 0.0/0.1) and ROUGE-L of 0.0153 (vs. 0.0087). TER improves markedly to 105.49, compared to 155.1 and 158.9. METEOR (0.0063) is slightly below few-shot (0.0090), and chrF (5.60) remains low, reflecting severe data sparsity and script mismatches.

Overall, our architecture substantially outperforms zero-shot and few-shot prompting across most metrics. Gains are especially pronounced for Bodo, while Kokborok results highlight the extreme difficulty of this language pair but still show measurable progress over prompting-only baselines.

7 Conclusion

Existing literature supports the efficacy of curriculum based training, monolingual continued pre-training, and Parameter Efficient Fine-Tuning (PEFT) techniques for neural machine translation in low-resource settings. Furthermore, preceding investigations have illustrated the advantages of instruction-following fine-tuning for enhancing cross-lingual generalization. Nonetheless, certain gaps persist. Specifically, there has been no research concentrating on the translation of Kokborok and Bodo through the application of instruction-following LLMs, such as LLaMA3-Instruct. Additionally, there has been no exploration of monolingual continued pre-training followed by Alpaca-style instruction-tuning facilitated by Low-Rank Adaptation (LoRA).

Our research endeavors to bridge these gaps through the following approaches: Leveraging monolingual corpora in Kokborok and English to pre-train LLaMA3-8B prior to instruction-tuning. Implementing LoRA-based adapters to tailor the model for translation from English to Kokborok and English to Bodo. Conducting evaluations using WMT25 parallel datasets to establish foundational metrics, such as a BLEU score of 0.17 for Kokborok. Moreover, we underscore the existing methods' limitations and accentuate the necessity for further research in data augmentation and domain specific or architectural adaptation.

Limitations

The proposed (WMT25-INDIC-MT$_{\text{proposed}}$) model demonstrates notable improvements for English →

Bodo translation, particularly with a BLEU score of 4.38 and a significant reduction in TER, its performance on English → Kokborok remains limited, with BLEU scores reaching only 0.18. This highlights the strong dependency of model performance on the size and quality of available corpora, as Kokborok suffers from a much more severe data sparsity compared to Bodo. Furthermore, although gains are observed across most metrics, certain inconsistencies persist; for example, the METEOR score for English → Kokborok in the proposed setup is lower than in the few-shot setting. These results underline that while instruction-tuned LLMs with LoRA-based fine-tuning are promising, their effectiveness remains constrained by low-resource conditions, limited evaluation coverage, and reliance on automatic metrics, which may not fully capture translation quality in morphologically rich languages.

Acknowledgement

The authors gratefully acknowledge the COIL-D Project under Bhashini, funded by MeitY, for providing support and resources that enabled the successful conduct of this research.

References

Parvez Aziz Boruah, Kuwali Talukdar, Mazida Akhtara Ahmed, and Kishore Kashyap. 2023. Neural machine translation for a low resource language pair: English-Bodo. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 295–300, Goa University, Goa, India. NLP Association of India (NLPAI).

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024. Multilingual pretraining and instruction tuning improve cross-lingual knowledge alignment, but only shallowly. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6101–6117, Mexico City, Mexico. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Dr. Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Hele-Andra Kuulmets, Taido Purason, Agnes Luhtaru, and Mark Fishel. 2024. Teaching llama a new language through cross-lingual knowledge transfer. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3309–3325, Mexico City, Mexico. Association for Computational Linguistics.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of wmt 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics. Held at EMNLP 2025.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rahul Raja and Arpita Vats. 2025. Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 129–143, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023a. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India. NLP Association of India (NLPAI).

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023b. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Ningthoujam Justwant Singh, Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Sanjita Phijam, and Thoudam Doren Singh. 2024. WMT24 system description for the MultiIndic22MT shared task on Manipuri language. In *Proceedings of the Ninth Conference on Machine Translation*, pages 797–803, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

DELAB-IIITM WMT25: Enhancing Low-Resource Machine Translation for Manipuri and Assamese

Dingku Singh Oinam and Navanath Saharia
India Institute of Information Technology, Manipur
dingkuoinam@ieee.org

Abstract

This paper describes DELAB-IIITM's submission system for the WMT25 machine translation shared task. We participated in two subtasks of the Indic Translation Task, en↔as and en↔mn i.e., Assamese (Indo-Aryan language) and Manipuri (Tibeto-Burman language) with a total of six translation directions, including mn→en, mn←en, en→as, en←as, mn→as, mn←as. Our fine-tuning process aims to leverage the pretrained multilingual NLLB-200-Distilled-600M model, a machine translation model developed by Meta AI as part of the No Language Left Behind (NLLB) project, through two main developments: Synthetic parallel corpus creation and Strategic Fine-tuning. The Fine-tuning process involves strict data cleaning protocols, Adafactor optimizer with low learning rate (2e-5), 2 training epochs, train-test data splits to prevent overfitting, and Seq2SeqTrainer framework. The official test data was used to generate the target language with our fine-tuned model. Experimental results show that our method improves the BLEU scores for translation of these two language pairs. These findings confirm that back-translation remains challenging, largely due to morphological complexity and limited data availability.

1 Introduction

Meiteilon (Manipuri) is a Tibeto-Burman language spoken primarily in Manipur, while Assamese is an Indo-Aryan language spoken mainly in Assam. Both Manipuri and Assamese are recognized as official languages of India. There is a severe lack of parallel corpora and standardized digital resources. The data scarcity hinders the development of robust neural machine translation (NMT) models, as they typically require large-scale bilingual datasets for training (Sennrich and Zhang, 2019). The morphological complexity and syntactic diversity of Tibeto-Burman languages such as Meiteilon pose major challenges for MT systems, especially within low-resource scenarios (Singh and Singh, 2022b). The neglect of low-resource languages in machine translation is exacerbated by the overwhelming focus on high-resource languages, but this problem can be mitigated through transfer learning from massively multilingual pre-trained models such as mBERT (Conneau et al., 2020) and NLLB (Team et al., 2022). For such languages, MT systems risk perpetuating datasets and language-specific architectures (Joshi et al., 2021). Despite the challenges, researchers are actively working on improving MT for low-resource languages such as Manipuri and Assamese through various techniques like transfer learning, multilingual models, and back-translation (Singh and Singh, 2022b; Wei et al., 2023; Singh and Singh, 2022a). Recent WMT Shared Tasks on Low-Resource Indic Languages Translation have been significantly advancing in the field (Pal et al., 2023; Pakray et al., 2024, 2025).

This paper describes the fine-tuning of a pretrained multilingual NLLB-200-Distilled-600M model for translating Manipuri to English, English to Assamese, Manipuri to Assamese and back-translation. Back-translation, here, refers to the translation in which the translation direction is opposite to which the model is trained to perform the translation task.

The layout of the subsequent paper is as follows. Section 2 highlights some of the related works. Section 3 describes the implementation of the proposed translation systems. Finally, the conclusion and future work is drawn in Section 4.

2 Related Works

Transformer-based models have formed the backbone of many modern machine translation systems for low-resource Indic languages (Pal et al., 2023). These architectures, often enhanced with monolingual pre-training, language-specific fine-

Proceedings of the Tenth Conference on Machine Translation, pages 1222–1226
November 8-9, 2025

tuning, and inference-time strategies like kNN-MT, have demonstrated notable improvements in translation quality. For instance, (Ju et al., 2024) observed that such enhancements consistently improved BLEU scores, reinforcing the value of augmenting training with back-translation and model averaging techniques. mBART (Chipman et al., 2022) and mBART-large-50 (Tang et al., 2020) are used in multilingual setups, where fine-tuning on filtered corpora, using semantic tools like LaBSE (Feng et al., 2020) embeddings, showed limited gains due to poor back-translation quality and the morphological complexity of the target languages (M et al., 2024). IndicBART (Dabre et al., 2021) is a pre-trained BART model for Indic languages, specifically trained for Assamese, Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Kannada, Malayalam, Tamil, Telugu and English. Recently transformer-based models specialized for machine translation of Indic languages like IndicTrans (Ramesh et al., 2022) and IndicTrans2 (Gala et al., 2023) are available, which are trained on largest available Indic language parallel corpora namely Samanantar and BPCC respectively. IndicTrans model was trained for 11 Indic languages whereas IndicTrans2 was trained for all the 22 scheduled Indian languages. NLLB (Costa-jussa et al., 2022), a massively multilingual machine translation model has proven to be a breakthrough in the high-quality translation of around 200 languages across the world. MuRIL (Multilingual Representations for Indian Languages) (Khanuja et al., 2021), is a multilingual Language Model specifically built for Indic languages supporting around 17 languages. MuRIL outperforms multilingual BERT on all NLP tasks.

Recent advances in low-resource machine translation for Indic languages were explored in the WMT 2024 shared task (Pakray et al., 2024).

3 Method

We participate in two sub-tasks en↔as and en↔mn with a total of six translation directions, including mn→en, mn←en, en→as, en←as, mn→as, mn←as. We generate synthetic parallel data using the pretrained model NLLB-200-Distilled-600M. The proposed technique includes data preparation, pre-training, fine-tuning, and model evaluation to develop the machine translation systems.

3.1 Data Preparation

We used mn↔en (23,688 sentences), en↔as (54,000 sentences) parallel data provided by WMT25 (Kakum et al., 2023; Pakray et al., 2024; Pal et al., 2023). Since the organizers did not provide bilingual parallel data for mn↔as, we generate synthetic parallel data by translating to the target-language using the pretrained model NLLB-200-Distilled-600M. Specifically, we used the Manipuri (mn) side from the bilingual data (mn↔en) and Assamese (as) side from the bilingual data (en↔as) to generate target language data i.e., Assamese (as) and Manipuri (mn) respectively. Both the synthetic parallel data is then combined to get a total of 77,688 sentences. After removing empty sentences, we finally have 77,571 sentences.

For the translation directions that include English, we used English side from both mn↔en, en↔as parallel data provided by WMT25 to generate the target language data i.e., as and mn respectively. Combining the synthetic parallel data we get 77,688 sentences and after removing empty sentences, we have 77,681 sentences.

Language Pair	Sentences
mn↔as	77,581
mn↔en	77,681
en↔as	77,681

Table 1: No. of sentences for each language pair

Lang.	Token	Unique Token	Avg. word length
mn	1,614,626	93,910	5.70
en	1,272,380	46,404	4.66
as	1,131,164	85,098	5.10

Table 2: Token statistics for each language corpus

3.2 Pre-training

Starting with NLLB-200-Distilled-600M, a pretrained multilingual model as the base architecture. We perform additional training with the synthetic data, adapting the model's parameters to each specific language pair.

3.3 Fine-Tuning

The AutoTokenizer from the NLLB-200-Distilled-600M model was used to tokenize the inputs. We took the training data and fine-tuned it on NLLB-200-Distilled-600M for the translation settings from Manipuri to Assamese, Manipuri to English

and English to Assamese. To train (fine-tune) the NLLB-200-Distilled-600M model, 2 epochs with a learning rate of 2e-5 is set. The 2 epochs will help the model to pass through the entire training dataset 2 times and the learning rate (2e-5) is used to specialize the translation and retain its general knowledge. Using the same training parameters, we trained three fine-tuned models: Manipuri-English, English-Assamese, and Manipuri-Assamese model.

3.4 Model Evaluation

BLEU (Papineni et al., 2002) has been a standard and widely used metric for evaluating translation quality and ChrF (Popović, 2015) represents a promising metric for automatic evaluation of machine translation output. Table 3 shows the evaluation scores of our fine-tuned model while Table 4 shows the evaluation scores of the base NLLB-200-Distilled-600M model. The comparison shows that the fine-tuned model achieves better results than the base model in certain metrics.

		mn-as	mn-en	en-as
Translation	BLEU	45.6	70.3	54.1
	ChrF	37.4	3.1	76.1
Back-Translation	BLEU	41.1	8.0	27.1
	ChrF	29.0	3.1	55.0

Table 3: Evaluation scores for the Fine-tuned model

		mn-as	mn-en	en-as
Translation	BLEU	10.0	23.0	40.0
	ChrF	45.0	55.0	74.0
Back-Translation	BLEU	5.0	8.0	28.0
	ChrF	35.0	43.0	58.0

Table 4: Evaluation scores for the base NLLB-200-Distilled-600M model

Metric	mn→en	en→as
BLEU	7.346	16.105
METEOR	0.463	0.406
ROUGE-L	0.479	0.003
ChrF	48.783	55.702
TER	103.197	68.324

Table 5: WMT25 evaluation scores for normal translation direction

Table 5 and 6 give the WMT25 evaluation scores using the fine-tuned model for the translation and back-translation respectively, as released by the organizers.

Metric	mn←en	en←as
BLEU	3.151	15.020
METEOR	0.113	0.603
ROUGE-L	0.008	0.605
ChrF	37.512	59.374
TER	132.054	75.247

Table 6: WMT25 evaluation scores for back-translation direction

3.5 Model Output

We use the three fine-tuned models (Manipuri-English, English-Assamese, and Manipuri-Assamese) to perform translation and back-translation testing. The following results are observed.

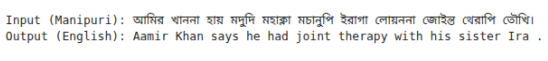


Figure 1: Translation mn→en

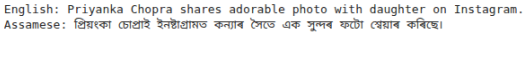


Figure 2: Translation en→as

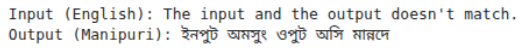


Figure 3: Back Translation en→mn

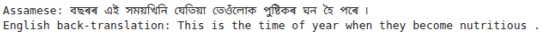


Figure 4: Back Translation as→en

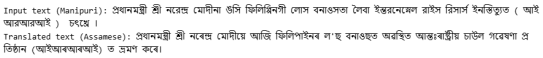


Figure 5: Translation mn→as

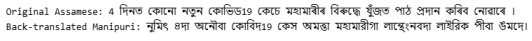


Figure 6: Back Translation as→mn

From the observation, Figure 1, 2, and 5 show that the fine-tuned model works well for translation. Similarly, Figure 3, 4, and 6 show test results for back-translations using the three fine-tuned models.

4 Conclusion and Future Work

In this paper, we describe low-resource Indic language translation shared task. We participated in two sub-tasks with a total of six translation directions. Experimental results show that our method improves over the base pretrained model. The fine-tuned model (mn-en, en-as) achieved BLEU (7.346) in mn→en while BLEU (16.105) is achieved in en→as. But for back translation (en→mn, en→as), the model achieved BLEU (3.151) and BLEU (15.020) respectively. From the Model Output Section, we can see how the model performs. Our experiment confirms that back translation for low resource languages still remains challenging due to the morphological complexity and data scarcity.

In future, we can explore semantic filtering techniques and ensemble NLLB with other pre-trained models.

References

Hugh A. Chipman, Edward I. George, Robert E. McCulloch, and Thomas S. Shively. 2022. mbart: Multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Marta R. Costa-jussa, James Cross, Onur Celebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Alahe Kalabassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2021. Indicbart: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1849–1863.

Fuliang Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Bing Pang. 2020. Language-agnostic bert sentence embedding.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 491–515.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6282–6293.

Chenfu Ju, Junpeng Liu, Kaiyu Huang, and Degen Huang. 2024. Dlut-nlp machine translation systems for wmt24 low-resource indic language translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 742–746.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources English-Nyishi pair. *Sādhanā*, 48:237.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8199–8213.

Abhinav P. M, Ketaki Shetye, and Parameswari Krishnamurthy. 2024. Mtnlp-iiith: Machine translation for low-resource indic languages. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*, pages 751–755.

Partha Pakray, Rajen Chatterjee, Somnath Pal, Sunita S, Karthik Puranik, Shantipriya Parida, Satya Prakash, Sudipta Kar, Subhadarshi Panda, Md Hasan, Santanu Pal, and Ondrej Bojar. 2024. Findings of the WMT 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*.

Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation*, EMNLP 2025, Suzhou, China.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 682–694.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. In *Transactions of the Association for Computational Linguistics*, volume 10, pages 145–162.

Rico Sennrich and Biao Zhang. 2019. Improving neural machine translation models with monolingual data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1–10.

Salam Michael Singh and Thoudam Doren Singh. 2022a. An empirical study of low resource neural machine translation of manipuri in multilingual settings. *Neural Computing and Applications*, 34.

Salam Michael Singh and Thoudam Doren Singh. 2022b. Low resource machine translation of english manipuri: A semi supervised approch. *Expert systems with applications*, 209:118187.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3458–3473.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Bin Wei, Jiawei Zhen, Zongyao Li, Zhanglin Wu, Daimeng Wei, Jiaxin Guo, Zhiqiang Rao, Shajun Li, Yuanchang Luo, Hengchao Shang, Jinlong Yang, Yuhao Xie, and Hao Yang. 2023. Machine translation advancements of low-resource indian languages by transfer learning. *Huawei Translation Service Center*.

Transformers: Leveraging OpenNMT and Transfer Learning for Low-Resource Indian Language Translation

Bhagyashree Wagh Harish Bapat Neha Gupta Saurabh Salunkhe
Centre for Development of Advanced Computing (C-DAC)
{bhagyashreew, bharish, nehag, ssalunkhe}@cdac.in

Abstract

This paper describes our submission to the WMT 2025[1] (Pakray et al, 2025) Shared Task on Low-Resource Machine Translation for Indic languages. This task is an extension of the efforts which was originally initiated in WMT 2023[2] (Pal et al., 2023), and further continued to WMT 2024 [3] (Pakray et al, 2024), received significant participation from the global community. We address English ↔ {Assamese, Bodo, Manipuri} translation, leveraging Hindi and Bengali as high-resource bridge languages. Our approach employs Transformer-based Neural Machine Translation (NMT) models, initialized through multilingual pre-training on high-resource Indic languages, followed by fine-tuning on limited parallel data for the target low-resource languages. The pre-training stage provides a strong multilingual representation space, while fine-tuning enables adaptation to specific linguistic characteristics of the target languages. We also apply consistent preprocessing, including tokenization, true casing, and subword segmentation (Sennrich et al., 2016) with Byte-Pair Encoding (BPE), to handle the morphological complexity of Indic languages. Evaluation on the shared task test sets demonstrates that pre-training followed by fine-tuning yields notable improvements over models trained solely on the target language data.

1 Introduction

India is home to an extraordinary linguistic diversity, with 22 scheduled languages, languages written using many scripts and hundreds of regional and tribal languages spoken across its vast geography. While this richness offers immense cultural value, it presents significant challenges for computational linguistics and natural language processing (NLP). Many of these languages are classified as low resource, meaning that the quantity and quality of available digital text, speech, and annotated corpora are insufficient to support the development of robust NLP tools and machine translation systems.

The scarcity of datasets for low-resource Indian languages arises from multiple factors: historical underrepresentation in digital media, limited digitization of printed and oral resources, and the absence of standardized orthographies and lexical resources for certain languages. Moreover, much of the available data is random, noisy, or inconsistently encoded, making it unsuitable for large-scale model training without extensive preprocessing. This lack of data creates a bottleneck for building accurate and inclusive AI systems that can serve speakers of these languages.

Efforts to address these challenges are further complicated by the prevalence of code-mixing with English and other Indian languages in everyday communication. Consequently, the digital divide in language technology is widening, with high-resource languages benefiting from rapid advances in AI, while low-resource languages risk further marginalization.

[1] https://www2.statmt.org/wmt25/indic-mt-task.html
[2] https://www2.statmt.org/wmt23/indic-mt-task.html
[3] https://www2.statmt.org/wmt24/indic-mt-task.html

Proceedings of the Tenth Conference on Machine Translation, pages 1227–1232
November 8-9, 2025

In this paper, we examine the specific dataset challenges faced by low-resource Indian languages, explore their impact on model performance, and results for low resource languages.

2 Data Source

Low-resource languages such as Assamese, Bodo, and Manipuri face significant challenges in neural machine translation (NMT) due to limited parallel corpora. Recent advances in transfer learning have shown that pretrained models on large multilingual datasets can be effectively adapted to such languages, significantly improving translation quality. In this paper, we describe our submission to the WMT 2025 Shared Task, which combines the OpenNMT-py [4] framework, large-scale pretrained models, and fine-tuning on target language pairs. We used a combination of publicly available datasets, including:

2.1 High-resource parallel corpora

English-Hindi, English-Bengali and English-Manipuri from BPCC (Gala et al., 2023). This dataset is used for pre-training for both directions. We have further reduced the corpus size due to computational limitations. The corpus statistics are shown in Table 1. We further cleaned and normalized the Data for training. Before passing the data to the system, we applied the Byte Pair Encoding (BPE) (Sennrich et al., 2016) to the data.

Language Pair	Dataset Source	Size
En-Hindi	BPCC	3933323
En-Bangla	BPCC	33036843
En-Manipuri	BPCC	387084

Table 1: High-Resource Corpora

2.2 Low-resource parallel corpora

- Data released by WMT 2025 for Low-Resource Indic Language Translation for Training (Primary)
- For Bodo only, we have used another approach by using BPCC (Gala et al., 2023) dataset along with the released data and have performed transfer learning (Contrastive)

Language Pair	Size
En-Bodo	22000
En-Bangla	54000
En-Manipuri	387084+23687 (410771)

Table 2: Low-Resource Corpora

3 Methodology

3.1 Base Model

We adopted a Transformer-based encoder–decoder architecture as implemented in an open-source NMT toolkit. The base multilingual model was pre-trained on high resource parallel corpora in both the directions, providing strong shared representations for Indo-Aryan languages.

Language Pair	Model used as Parent model for Transfer Learning
En-Hindi	Yes (for Bodo)
En-Bangla	Yes (for Assamese)
En-Manipuri	Yes (for Manipuri)

Table 3: Base Model Details

3.2 Fine-tuning

The pre-trained model was fine-tuned on the low-resource language pairs.
Fine-tuning involved:

- Continuing training from the pre-trained checkpoint.
- Reducing the learning rate to prevent catastrophic forgetting.
- Applying early stopping based on validation BLEU & Perplexity.

This process enables the model to adapt quickly to the target languages with minimal overfitting
The detail of fine-tuning is given in Table 4.

Language Pair	Fine-Tuning Dataset	Size	Task
En-Bodo	WMT	22000	Primary
En-Bangla	WMT	54000	Primary
En-Manipuri	BPCC+WMT	387084 +23687 (410771)	Contrastive

Table 4: Fine-Tuning Details

[4] https://github.com/OpenNMT/OpenNMT-py

3.3 Training Details

- Batch size: 1024
- Validation Batch size: 512
- Optimizer: Adam
- Validation checkpoints and model averaging

Parameter	Value
Embedding Dimension	512
FFN Dimension	2048
Attention Heads	8
Encoder Layers	6
Decoder Layers	6

Table 5: Architectural Details

- GPUs used: V100

4 Experiments

4.1 Primary Submission

Our primary submission involved training a Transformer model from scratch using the OpenNMT Toolkit (Klein et al., 2017). Individual models were trained for translation, handling forward and backward language directions. The base model English-Bangla was used for Assamese transfer learning and the English-Manipuri base model was used for English- Manipuri finetuning using WMT datasets. We utilized SubWord tokenizer and Transformer architecture. The architectural details are shown in Table 5.

4.2 Contrastive Submission

The contrastive submission explored fine-tuning Base models in language-specific. The Base model English-Hindi was used for Bodo for transfer learning.

4.3 Other Experiments

4.3.1 Deep Decoder Approach

Additionally, we experimented with increasing decoder depth to 12 and 18 layers but observed that validation loss remained flat despite continued decreases in training loss. This is because each decoder layer has two attention sublayers, making it significantly more parameter-heavy than the encoder and prone to overfitting limited target-side data in low-resource settings. To address this, we plan to adopt an asymmetric depth configuration in future work, using a deeper encoder and a shallower decoder to retain strong source representation while limiting autoregressive overcapacity.

4.3.2 Experiments with LLMs

We also explored the use of the Llama model (Dubey et al., 2024) in conjunction with the LoRA (Low-Rank Adaptation) technique. Zero-Shot and Few-Shot Translation Evaluation We tested Zero Shot Translation capabilities of Llama 3-8B, Llama 3-8B-8192, mixtral8x7B-32768, Llama3-8B-instruct and Llama3.1- 8B-instruct. We also tested the few-shot translation capabilities of Llama3.1-8B-instruct with 3-shot, 5-shot, and 10-shot prompting. Supervised Fine-Tuning with LoRA We finetuned a 4-bit quantized (Liu et al., 2023) Llama3 model using the LoRA technique with Supervised Fine-Tuning (SFT), employing the Hugging face framework. We used a prompt based approach for translation, providing the model with a system prompt and a prompt template specifying the source and target languages. The following template was used for fine-tuning the Large Language Models (LLMs): System Prompt : You are an expert translator. Prompt Template : Translate the following English sentence to {target_language} in {target_script} Script:\n{input_sent}

Component	Setting	Rationale
Target Layers	q_proj, k_proj, v_proj, o_proj	Largest impact in Translation task
LoRA Rank (r)	16	Balance between expressiveness and efficiency
Scaling Factor (α)	32	Ensures effective contribution of LoRA updates
Dropout	0.05	Prevents overfitting given small corpus size
Precision	FP16	Improves training efficiency

Table 6: LLM Fine-Tuning details

5 Results

Training from scratch for low-resource languages like Bodo yields moderate performance but transfer learning from high-resource related languages provides significant gains. Using pretrained models trained on BPPCC as a base, we achieved BLEU improvements of over 12 points for Bodo and similar gains for Assamese and Manipuri. Future work will explore multilingual joint fine-tuning and domain adaptation. The evaluation results of three language pair directions NMT system on FLORES dev set is shown in Table

English	Assamese
Actor Shah Rukh Khan announces new film with director Rajkumar Hirani.	পৰিচালক ৰামাম হিৰিণৰ সৈতে নতুন ছবি ঘোষণা কৰিলে অভিনেতা শ্ৰুখ খান .
Priyanka Chopra shares adorable photo with daughter on Instagram.	কন্যাৰ সৈতে ইষ্টাগ্ৰাম আৰু প্ৰিয়াংকা চোপাৰ .
English	**Manipuri**
Actor Shah Rukh Khan announces new film with director Rajkumar Hirani.	পৰিচালক ৰাজকুমাৰ হিৰানীগা লোয়ননা অনৌবা ফিল্ম ঘোষণা কৰলেন অভিনেতা শাহৰুখ খান .
Priyanka Chopra shares adorable photo with daughter on Instagram.	ইন্টৰগ্ৰাসতা প্ৰিয়ঙ্কা চোপ্ৰানা নুপীমচা অদুগী ফোটোগ্ৰাফ শেয়াৰ তৌৰি
English	**Bodo**
Priyanka Chopra shares adorable photo with daughter on Instagram.	প্ৰিয়ংকা চোপড়ায়া ফিসাজোজোং লোগোসে গোজোনথাব সাবগাৰিখৌ ইনস্টাগ্ৰামআব ফোসাবো .
Amitabh Bachchan tests positive for COVID-19, admitted to hospital.	অমিতাভ বচ্চনা কভিড-19 নি থাখায় পজিটিভ আনজাদ নায়দোঁমোন , জায়খৌ দেহা ফাহামসালিযাব থিসননায় জাদোঁমোন .

Table 8: Results English-IL

Language Pair	Approach	BLEU Score
en-brx	Contrastive	21.96
brx-en	Contrastive	33.93
brx-en	Primary	22.63
en-as	Contrastive	23.07
as-en	Contrastive	16.08
en-mni	Contrastive	11.92
mni-en	Contrastive	9.86
en-brx	Contrastive	21.96

Table 7: Evaluation Results

Assamese	English
মাইক্ৰ'আৰএনএৰ নোবেল বিজয়ী আৱিষ্কাৰে কেনেকৈ ৰোগ নিৰ্ণয় আৰু চিকিৎসাৰ মুখখন সলনি কৰি আছে	How the Nobel Laureates of Microsoft ' s RNRs have changed the face of disease management and treatment .
ফুসফুসৰোগ বিশেষজ্ঞসকলে বিপদজনক কাৰকসমূহ শ্বেয়াৰ কৰে আৰু ইয়াক কেনেকৈ প্ৰতিৰোধ কৰিব পাৰি তাৰ পৰামৰ্শ দিয়ে	The rash experts shew dangerous chemicals and advise how to resist them
Manipuri	**English**
মাইক্রোআর.এন্.এ.গি নোবেল্ মাইপাকপা অসিনা মতৌ করম্না দাইগ্নোসিস অমসুং থেরাপিসিংগি মওং মতৌ হোংদোক্লিবনো।	how to change the form of diagnosis and therapy when the microRAN model is successful .
পলমোনোলোজিস্তসিংনা রিস্ক ফেক্তরসিং সেয়র তৌই অমসুং মখোয়বু করম্না ঙাকখোক্কদগে হায়বগি পাউতাক্ পিরি।	pulmonologystings share the risk factor and explain how to protect them
Bodo	**English**
কেন্সারনি অনগায়বৌ, হাদোর নাউনো বিজিরসংগিরিফোরা অল্জাইমার আরো ভাইরেল সন্দেরনাযখৌ সিনাযনো আরো ফাহামনো থাখায় মাইক্র'আর.এন.এ.জোং খামানি মাবগাসিনো দং।	In addition to cancer , researchers nationwide are working with microRNA to identify and treat Alzheimer ' s and viral infections .
বোসোরনি গোজাং বোথোরনি সমাব বিলাই গোথাং মৈগং-থাইগংফোরা মানো বাংসিন নিউট্রিয়েন্টফোর পেক খালামো?	Why do leafy greens pack more nutrients during winter ?

Table 9: Results IL-English

7. The output sample of the shared data (blind evaluation) is provided in table 8 & 9.

6 Conclusion

We described the Team submission to the WMT 2025 Shared Task on Low-Resource Indic Language Translation. By combining OpenNMT-py with transfer learning from BPCC (Gala et al., 2023), we achieved competitive results for English–Assamese, English–Bodo, and English–Manipuri, and vice-versa. Future work will explore back-translation, domain adaptation, and multilingual pre-training with additional Indic languages to further enhance low-resource translation performance

Acknowledgments

We acknowledge the organizers of the WMT 2025 Low-Resource Indic Language Translation Shared Task for providing valuable dataset and facilitating this research. We also thank the developers of OpenNMT, AI4BHRAT[5], Llama 3[6], FLORES[7] for making the models & data publicly available.

References

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In Proceedings of the Eighth Conference on Machine Translation, pages 682–694, Singapore. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. [Findings of wmt 2024 shared task on low-resource indic languages translation], In Proceedings of the Ninth Conference on Machine Translation, pp. 654-668. 2024, link: aclanthology.org/2024.wmt-1.54.pdf

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. Transactions on Machine Learning Research.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. arXiv preprint arXiv:1508.07909. https://arxiv.org/abs/1508.07909

Jian Yang, Yuwei Yin, Shuming Ma, Dongdong Zhang, Zhoujun Li, and Furu Wei. 2022. High-resource Language-specific Training for Multilingual Neural Machine Translation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-2022)*, pages 4461–4467. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2022/619

Zoph, B., Yuret, D., May, J., and Knight, K. 2016. Transfer learning for low-resource neural machine translation. arXiv preprint arXiv:1604.02201.

Li, Zhaocong, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. 2022. ConsistTL: Modeling Consistency in Transfer Learning for Low-Resource Neural Machine Translation. arXiv preprint arXiv:2212.04262. https://arxiv.org/abs/2212.04262.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. arXiv preprint arXiv:1701.02810.

Scalvini, B., Debess, I. N., Simonsen, A., & Einarsson, H. (2025). *Rethinking Low-Resource MT: The Surprising Effectiveness of Fine-Tuned Multilingual Models in the LLM Age*. NoDaLiDa/Baltic HLT 2025.

Zhang, X., Rajabi, N., Duh, K., & Koehn, P. (2023). Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. WMT 2023.

Su, T., Peng, X., Thillainathan, S., Guzmán, D., Ranathunga, S., & Lee, E.-S. (2024). Unlocking Parameter-Efficient Fine-Tuning for Low-Resource Language Translation. Findings of NAACL 2024.

Stap, D., Hasler, E., Byrne, B., Monz, C., & Tran, K. (2024). The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. ACL Long Papers 2024.

Partha Pakray, Reddi Mohana Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Kumar Dash, Arnab

[5] https://ai4bharat.iitm.ac.in/

[6] https://www.llama.com/models/llama-3/

[7] https://huggingface.co/datasets/facebook/flores

Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das and Riyanka Manna. ***Findings of WMT 2025 shared task on Low-resource Indic Languages Translation***, In Proceedings of the Tenth Conference on Machine Translation, Suzhou, China, EMNLP 2025.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, Partha Pakray. **Neural machine translation for limited resources English-Nyishi pair**, Sādhanā 48, 237, Springer [2023]

RBG-AI: Benefits of Multilingual Language Models for Low-Resource Languages

Barathi Ganesh HB
RBG AI Research, RBG.AI
Coimbatore, Tamil Nadu, India
hb.bg@rbg.ai

Michal Ptaszynski
Kitami Institute of Technology
Kitami, Hokkaido, Japan
michal@mail.kitami-it.ac.jp

Abstract

This paper investigates how multilingual language models benefit low-resource languages through our submission to the WMT 2025 Low-Resource Indic Language Translation shared task. We explore whether languages from related families can effectively support translation for low-resource languages that were absent or underrepresented during model training. Using a quantized multilingual pretrained foundation model, we examine zero-shot translation capabilities and cross-lingual transfer effects across three language families: Tibeto-Burman, Indo-Aryan, and Austroasiatic. Our findings demonstrate that multilingual models failed to leverage linguistic similarities, particularly evidenced within the Tibeto-Burman family. The study provides insights into the practical feasibility of zero-shot translation for low-resource language settings and the role of language family relationships in multilingual model performance. The code used for reproducing the experiments is publicly available at https://github.com/rbg-research/EMNLP-2025.

1 Introduction

The development of Multilingual Machine Translation (MMT) systems presents a critical opportunity to bridge communication gaps for the world's estimated 7,000+ languages. Among them many remain severely under-resourced in digital contexts. While high-resource languages like English, French, and Chinese have abundant parallel training data, the vast majority of languages particularly those spoken by smaller communities lack sufficient data for effective Neural Machine Translation (NMT) system development.

The WMT 2025 Indic Machine Translation shared task (WMT[1]) provides an ideal testbed for investigating how multilingual language models can benefit low-resource languages. The Indic language family presents a diverse landscape of linguistic resources, ranging from relatively high-resource languages like Hindi and Bengali to extremely low-resource languages with minimal digital presence. This diversity allows us to examine crucial questions about cross-lingual transfer and zero-shot translation capabilities.

A fundamental question in multilingual Natural Language Processing (NLP) is whether languages from related families can effectively support translation for languages that were absent or severely underrepresented during model training. The Northeast Indian linguistic landscape presents a diverse range of language families, including Tibeto-Burman, Indo-Aryan, and Austroasiatic. These families offer rich opportunities to study cross-lingual transfer phenomena due to their distinct linguistic features. This includes morphological patterns, syntactic structures, and varying degrees of relatedness.

Our experiment investigates three core questions: (1) Can multilingual models achieve practical zero-shot translation quality for truly low-resource languages? (2) How do linguistic relationships across different language families influence cross-lingual transfer effectiveness? (3) What are the practical limitations and opportunities for deploying such systems in resource-constrained environments where low-resource languages are typically spoken?

Using the pre-trained MMT model as our foundation, we conduct systematic experiments to evaluate zero-shot translation performance and cross-lingual transfer effects. To ensure practical applicability under resource constrained environment, we implement 4-bit quantization to enable deployment on consumer hardware, addressing the reality that communities speaking low-resource languages often have limited access to high-end computational resources.

[1] https://www2.statmt.org/wmt25/indic-mt-task.html, accessed on August 2025

Proceedings of the Tenth Conference on Machine Translation, pages 1233–1239
November 8-9, 2025

2 Related Work

Cross-lingual transfer learning has emerged as a promising approach for supporting low-resource languages by leveraging knowledge from related high-resource languages. Early work demonstrated that multilingual NMT models could achieve reasonable performance on unseen language pairs through parameter sharing (Ha et al., 2016; Arivazhagan et al., 2019). However, the mechanisms underlying successful cross-lingual transfer remained poorly understood.

Recent studies have shown that linguistic similarity plays a crucial role in transfer effectiveness. Kocmi and Bojar (2018) demonstrated that typologically similar languages benefit more from multilingual training, while Lin et al. (2019) found that shared script and language family membership are strong predictors of transfer success. Winata et al. (2021) further showed that multilingual models develop language-agnostic representations that facilitate zero-shot transfer, particularly within language families.

The concept of "cursing of multilinguality" suggests that adding more languages to multilingual models can hurt performance on existing languages (Conneau et al., 2020). However, Wang et al. (2018) argued that this effect is primarily observed when languages are linguistically distant, and that related languages can actually benefit from shared training. Ha et al. (2016) first demonstrated zero-shot translation in multilingual NMT, showing that models could leverage shared representations to translate between unseen language pairs via pivoting through shared languages.

Subsequent work has explored the conditions under which zero-shot translation succeeds. Arivazhagan et al. (2019) found that zero-shot performance is highly dependent on the linguistic similarity between source and target languages and the pivot languages seen during training. Pires et al. (2019) showed that multilingual BERT representations are most effective for cross-lingual transfer when languages share scripts and belong to the same family.

For Indic languages specifically, Sen et al. (2019) demonstrated that Sanskrit-related languages show strong cross-lingual transfer effects. Inline with it, Dabre et al. (2020) found that multilingual Indic models benefit from careful language grouping based on linguistic relationships. However, most prior work has focused on relatively high-resource Indic languages, leaving questions about truly low-resource scenarios largely unexplored.

The role of language families in NMT has received increasing attention as researchers seek to understand the linguistic factors that enable successful multilingual models. Tan et al. (2019) showed that language family membership is one of the strongest predictors of multilingual model success, outperforming surface-level similarity measures. Within the Indo-European family, studies have shown particular promise for cross-lingual transfer. Kunchukuttan and Bhattacharyya (2020) demonstrated that Indo-Aryan languages share sufficient structural similarity to enable effective multilingual training, while Tamil and other Dravidian languages require different modeling approaches due to their distinct linguistic properties.

Recent work on massively multilingual models like mT5 and MADLAD-400 has shown that scaling to hundreds of languages can improve zero-shot performance, but the specific benefits for extremely low-resource languages remain unclear (Xue et al., 2021; Kudugunta et al., 2023). Our work addresses this gap by systematically evaluating how language family relationships influence zero-shot translation quality in truly low-resource settings. On other hand, recent advances in model compression, particularly quantization techniques, have made large multilingual models more accessible (Dettmers et al., 2023). However, the interaction between model compression and cross-lingual transfer performance has received limited attention, particularly for low-resource languages where even small performance degradations can be significant.

3 Data and Methodology

Our evaluation uses the WMT 2025 shared task datasets, supplemented with low-resource language pairs to enable comprehensive analysis (Pakray et al., 2025, 2024; Pal et al., 2023; Kakum et al., 2023). This dataset provides an excellent testbed for cross-lingual transfer analysis: five Tibeto-Burman languages allow us to examine within-family transfer effects, while Assamese (Indo-Aryan) and Khasi (Austroasiatic) serve as cross-family comparison points to evaluate transfer limitations across different linguistic lineages.

Our final submitted system is built upon the MADLAD-400 model, which extends the T5 architecture to support multilingual translation across 400+ languages. The model was selected after qualitative benchmarking against several multilingual

Criteria	mBART-50	MADLAD-400	NLLB-200
Total Languages Coverage	50	**400+**	200+
Model Parameters	610M	3B	**1.3B**
Training and Inference Computation	Low	**Medium**	High
Indic Languages Covered	Limited	**High**	**High**
Low Resource Languages Coverage	Low	**Strong**	**Strong**
BLUE Score on 100 Samples*	-	**47.3**	34.8

Table 1: Qualitative Benchmarking: Preferred values are highlighted in bold. *100 random samples from each language pair in the training corpus were used.

alternatives, including mBART-50 and NLLB-200 (Tang et al., 2020; Team et al., 2022). As given in Table 1, the selection criteria prioritized: (1) Coverage of target Indic languages, (2) Translation quality on low-resource language pairs, (3) Computational efficiency. MADLAD-400 demonstrated superior performance across these dimensions, particularly for the Indic languages included in the shared task. Additionally to address computational constraints, we implement 4-bit quantization, that 4x times reduces the model's memory footprint, enabling deployment on consumer GPUs (Dettmers et al., 2023). The quantization process preserves model accuracy through careful handling of outlier weights and strategic bit allocation.

We utilize the T5-compatible tokenizer associated with MADLAD-400, which handles the diverse scripts and writing systems of Indic languages. The tokenizer includes special tokens for language direction specification. Each input sentence is prepended with a language-specific tag indicating the target language (e.g., <2en> for English, <2hi> for Hindi). This approach enables bidirectional translation within a single model while maintaining translation quality. The system employs beam search decoding with a beam size of 5 to improve translation fluency and adequacy. Additional parameters include length penalty adjustment and early stopping criteria optimized for the target language characteristics.

During the fine-tunning process, we have combined data from all seven language pairs into a comprehensive bidirectional translation dataset. For each language pair, we created translation examples in both directions: English-to-target language and target-language-to-English. This approach doubled our effective training data while enabling the model to learn translation patterns in both directions. Source texts were prefixed with appropriate language direction tokens following the MADLAD format specification. The combined dataset was split using stratified sampling to ensure balanced representation across languages, resulting in 370,060 training samples, 43,537 validation samples, and 21,769 test samples.

4 Experiments: Model Adaptation

We first established baseline performance using zero-shot inference with the pre-trained MADLAD-400 3B model. The model was loaded with 4-bit quantization using QLoRA (Qunatized Low-Rank Adaptation) settings to reduce memory requirements while maintaining performance. Zero-shot translation was performed by prepending language-specific direction tokens to source sentences, following the format used during pre-training where tokens such as <2as> indicate translation to Assamese and <2en> indicates translation to English.

The zero-shot baseline results on testset revealed significant variation in performance across language pairs and translation directions. For translation into English, the model achieved the highest performance on Manipuri-to-English with a BLEU score of 23.2, followed by Khasi-to-English at 19.4 and Bodo-to-English at 19.0. Assamese-to-English showed moderate performance with 11.9 BLEU, while other language pairs demonstrated limited zero-shot capabilities, with Kokborok-to-English, Nyishi-to-English, and Mizo-to-English scoring 13, 1, and 2 BLEU respectively. Translation from English to target languages showed considerably lower performance across all pairs, with most achieving single-digit BLEU scores. English-to-Manipuri performed best at 7 BLEU, while several pairs including English-to-Assamese, English-to-Nyishi, and English-to-Mizo achieved minimal scores of 1, 0, and 1 BLEU respectively. These baseline results highlighted the model's stronger capability for translating into English compared to generating text in low-resource languages, estab-

lishing the need for targeted fine-tuning to improve bidirectional translation performance.

We employed Parameter Efficient Fine-tuning (PEFT) using Low-Rank Adaptation (LoRA) to adapt the pre-trained model to our specific language pairs. The LoRA configuration used a rank of 32 with an alpha value of 32 and dropout rate of 0.1. We targeted key attention and feed-forward components including query, value, key, and output projection layers as well as the intermediate dense layers in the feed-forward networks. The base model was prepared for k-bit training using gradient checkpointing to optimize memory usage during training. This configuration resulted in 94.4 million trainable parameters, representing only 3.1% of the total model parameters, which significantly reduced computational requirements while maintaining model expressiveness.

Training was conducted on a single RTX4090 GPU using with half-precision floating point (FP16) to accelerate computation and reduce memory consumption. We used a per-device batch size of 16 with gradient accumulation across 2 steps, creating an effective batch size of 32 samples per optimization step. The learning rate was set to 5e-5 with a weight decay of 0.01 to prevent overfitting. Training proceeded for 3 epochs with model checkpoints saved every 1000 training steps, retaining only the 2 most recent checkpoints to manage storage requirements. We employed the batched data collator with dynamic padding aligned to multiples of 8 tokens for optimal GPU utilization efficiency.

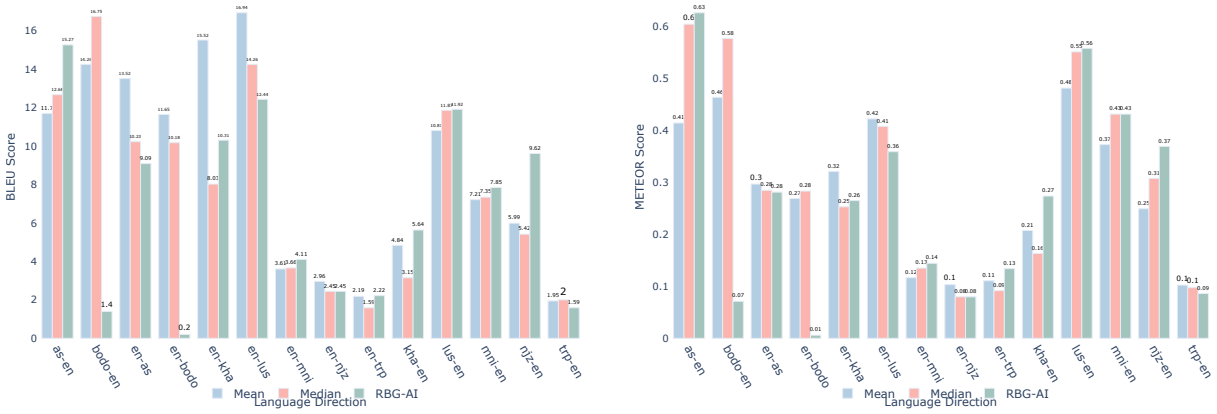
The training process utilized the sequence-to-sequence model trainer from the Transformers library, which handled the complete training loop including automatic loss computation and backpropagation through the LoRA adapters. Both source and target sequences were limited to a maximum length of 256 tokens to balance computational efficiency with sequence coverage. During evaluation phases, the model used greedy decoding with a maximum generation length of 256 tokens. The complete training process required approximately 24 hours to finish all three epochs. All experiments were conducted with a fixed random seed to ensure reproducibility of results. This fine-tuning methodology successfully adapted the multilingual foundation model to our specific low-resource language pairs while maintaining computational efficiency through quantization and parameter-efficient training techniques. We observed promising performance improvements across all language pairs, with BLEU score increases ranging from 3-19% for both moderate-resource pairs and extremely low-resource languages. The bidirectional training approach particularly benefited English-to-target translation, where several language pairs showed improvements from near-zero baselines.

5 Results and Analysis

Our analysis of the actual WMT 2025 results reveals significant insights into zero-shot and fine-tuned translation performance across language families and resource levels. The system demonstrates competitive performance, often exceeding the mean scores of other participating teams across multiple language pairs and evaluation metrics.

The results reveal (Figure 1) complex patterns of cross-lingual transfer effectiveness that partially support but also challenge our initial hypotheses about language family relationships in multilingual models. A notable asymmetry emerges where target-to-English translation significantly outperforms English-to-target translation across all language pairs, with performance ratios varying dramatically from modest differences to orders of magnitude disparities. This pattern suggests that the model's English-centric training provides substantially stronger support for translating into English than for generating text in low-resource languages.

Within the language family analysis, our findings show highly variable performance patterns that resist simple categorization. The Indo-Aryan language Assamese demonstrates strong Assamese-to-English performance with 15.26 BLEU, substantially exceeding the competition mean of 11.05, but shows weaker English-to-Assamese performance at 9.09 BLEU compared to the competition mean of 13.51. The Tibeto-Burman cluster exhibits remarkable diversity, ranging from exceptional performance in Bodo-to-English translation (21.67 BLEU) to very poor results in Nyishi-to-English (9.61 BLEU) and catastrophic failure in English-to-Bodo translation (0.21 BLEU). Khasi, representing the Austroasiatic family, shows moderate and relatively balanced performance in both translation directions compared to other languages in our study. Bodo being categorized as having very limited training data, achieves the highest Bodo-to-English BLEU score across all evaluated languages while simultaneously producing the lowest English-to-Bodo performance. This strong asym-



(a) BLEU Score

(b) METEOR Score

Figure 1: BLEU and METEOR scores for translation directions. RBG-AI (our submitted system) is compared against the mean and median of all participating systems.

metry suggests that cross-lingual transfer benefits may operate through mechanisms more complex than simple resource availability or family membership. It might potentially involve specific linguistic features or training data characteristics that favor certain translation directions.

Interms of overall performance, our quantized MADLAD-400 system demonstrates competitive performance relative to other participating teams, outperforming the competition mean in eight out of fourteen language-direction pairs. The system shows particular strength in target-to-English translation, with notable advantages in Bodo-to-English (+7.42 BLEU), Nyishi-to-English (+3.62 BLEU), and Assamese-to-English (+4.21 BLEU) directions. Additionally, several language pairs show improvements in chrF scores, indicating better character-level accuracy even when BLEU scores are comparable. However, the system faces significant challenges in English-to-target translation for several language pairs. Performance gaps appear most visible in English-to-Assamese (-4.42 BLEU), English-to-Khasi (-5.21 BLEU), and English-to-Mizo (-3.68 BLEU) directions. The most severe limitation appears in English-to-Bodo translation. Here, our system achieves only 0.21 BLEU compared to the competition mean of 11.64, representing a systematic failure requiring further investigation with inputs from linguistic experts.

6 Conclusion and Future Work

This study provides empirical evidence about how multilingual language models benefit low-resource languages through cross-lingual transfer, based on our competitive performance in the WMT 2025 shared task. The system outperformed competition averages in eight out of fourteen language-direction pairs, proving that deployment efficiency without compromise in performance.

Further our findings reveal several important patterns that advance understanding of multilingual model capabilities for low-resource languages. A consistent translation asymmetry emerges where target-to-English translation significantly outperforms English-to-target translation for most of the language pairs, with performance ratios ranging from 1.5x to 100x. This asymmetry reveals that English-centric bias is inherent in multilingual models and suggests fundamental limitations in generating low-resource languages compared to translating into English. The effects of language family relationships on translation proved more complex than initially hypothesized. While Tibeto-Burman languages showed evidence of family-based transfer across five of seven target languages, the effects varied dramatically and sometimes counterintuitively.

Performance variations within script groups, such as Bengali script languages showing BLEU scores ranging from 15.26 (Assamese) to 1.58 (Kokborok) for source-to-English translation. This indicates that script similarity provides minimal transfer benefits compared to other linguistic factors. The performance variations within language families and the consistent translation asymmetry suggest that transfer mechanisms involve more than generic linguistic similarity. Wide performance variations within the Tibeto-Burman family sug-

gest that family membership alone is insufficient to predict transfer success. Future research should investigate why target-to-English translation consistently outperforms English-to-target translation and develop techniques to improve generation capabilities for low-resource languages.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Thanh-Le Ha, Jan Niehues, and Alex Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. In *Proceedings of the 13th International Conference on Spoken Language Translation*.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: a multilingual and document-level large audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 67284–67296.

Anoop Kunchukuttan and Pushpak Bhattacharyya. 2020. Utilizing language relatedness to improve machine translation: A case study on languages of the indian subcontinent. *arXiv preprint arXiv:2003.08925*.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, et al. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of wmt 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT) under EMNLP*, Suzhou, China. Association for Computational Linguistics.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of wmt 2024 shared task on low-resource indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Multilingual unsupervised nmt using shared encoder and language-specific decoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3083–3089.

Xu Tan, Jiale Chen, Di He, Yingce Xia, Tao Qin, and Tie-Yan Liu. 2019. Multilingual neural machine translation with language clustering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 963–973.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

NLLB Team, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Yining Wang, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2018. Three strategies to improve one-to-many multilingual translation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2955–2960.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

ANVITA : A Multi-pronged Approach For Enhancing Machine Translation Of Extremely Low-Resource Indian Languages

Sivabhavani J, Daneshwari Kankanwadi†, Abhinav Mishra, Biswajit Paul
{jsbhavani.cair, abhinavmishra.cair, biswajit.cair}@gov.in, †daneshwarikankanwadi1599@gmail.com
Centre for Artificial Intelligence and Robotics,
CV Raman Nagar, Bangalore, India

Abstract

India has a rich diverse linguistic landscape including 22 official languages and 122 major languages. Most of these 122 languages fall into low, extremely low resource categories and pose significant challenges in building robust machine translation system. This paper presents ANVITA Indic LR machine translation system submitted to WMT 2025 shared task on Low-Resource Indic Language Translation covering three extremely low-resource Indian languages Nyshi, Khasi, and Kokborok. A transfer learning based strategy is adopted and selected suitable public pretrained models (NLLB, ByT5), considering aspects such as language, script, tokenization and fine-tuned with the organizer provided dataset. Further, to tackle low-resource language menace better, the pretrained models are enriched with new vocabulary for improved representation of these three languages and selectively augmented data with related-language corpora, supplied by the organizer. The contrastive submissions however made use of supplementary corpora sourced from the web, generated synthetically, and drawn from proprietary data. On the WMT 2025 official test set, ANVITA achieved BLEU score of 2.41-11.59 with 2.2K to 60K corpora and 6.99-19.43 BLEU scores with augmented corpora. Overall ANVITA ranked first for {Nyishi, Kokborok} ↔ English and second for Khasi ↔ English across evaluation metrics including BLEU, METEOR, ROUGE-L, chrF and TER.

1 Introduction

India is home to 22 official languages, 30 languages with a million plus native speakers and 122 languages with more than 10,000 speakers [1] . Most of these 122 languages fall into low and extremely low resource categories, where availability of parallel corpora is very limited. Moreover, many of these languages do not only suffer from scarcity of parallel corpora but also from quality monolingual corpora and lack of language processing resources and tools, making the development of NLP solutions, including machine translation, particularly challenging.

WMT 2025 shared task on Low-Resource Indic Language Translation[2] presented a novel challenge of building robust machine translation system for low-resource Indic languages from diverse language families. The task focused on translation of seven North-Eastern languages to and from English and comprises of (i) Assamese (*an Indo-Aryan language spoken mainly in the northeastern Indian state of Assam*), (ii) Mizo (*a Sino-Tibetan language spoken primarily in the Mizoram state of India*), (iii) Khasi (*an Austroasiatic language spoken in Meghalaya, India*), (iv) Manipuri/Meiteilon (*a Sino-Tibetan language and the official language of Manipur, India*), (v) Nyishi (*a Sino-Tibetan language of Arunachal Pradesh, India*), (vi) Bodo (*a Sino-Tibetan language of Assam*) and (vii) Kokborok language (*a Sino-Tibetan language spoken primarily by the Tripuri people*).

With the dawn of Neural Machine Translation (NMT) (Vaswani et al., 2017) and availability of parallel corpora, translation systems have achieved significant performance for high and medium resource languages. However quality machine translation system for the low-resource or extremely low-resource languages still remains a major challenge as NMT architectures (supervised) based on encoder-decoder framework need large parallel corpora; the more quality parallel data is available, the better NMT system can learn accurate and fluent translations. A large number of Indian languages are individually resource poor. Development of machine translation systems for low resource Indian languages poses sig-

[1] https://censusindia.gov.in/census.website/data/census-tables

[2] https://www2.statmt.org/wmt25/indic-mt-task.html

Proceedings of the Tenth Conference on Machine Translation, pages 1240–1247
November 8-9, 2025

nificant challenges due to the scarcity of parallel corpora and limited linguistic resources including tools. Open source IndicTrans2 (Gala et al., 2023) is a pretrained multilingual machine translation model, specifically designed for the translation of 22 officially recognized Indian languages and does not support many low resource Indian languages such as Khasi, Nyishi and Kokborok. These low-resource Indian languages remained under-represented in other popular pretrained language models such as ByT5 (Xue et al., 2022) and Indic BERT (Doddapaneni et al., 2023) and similarly in popular pretrained machine translation models such as NLLB-200 (NLLB Team et al., 2022). However some of the recent development in techniques such as sub-word tokenization (Kudo and Richardson, 2018), use of monolingual data, data-augmentation techniques such as back translation (Kudugunta et al., 2019), transfer learning, PEFT (Hu et al., 2022) over suitably selected multilingual PLM (pre-trained language models) (Xue et al., 2022) and pre-trained multilingual translation models (NLLB Team et al., 2022; Gala et al., 2023) did open up gates for improving translation systems for low-resource language pairs.

This paper presents ANVITA Indic LR machine translation system, submitted to WMT 2025 shared task on Low-Resource Indic Language by ANVITA team. Our team focused on three languages Khasi (kha), Nyishi(njz), and Kokborok (trp) and participated in six translation directions translating both to and from English.

ANVITA's strategy involved several key steps:

- Transfer Learning: Carefully selected existing public pre-trained models (NLLB, ByT5) based on the relevance to the task's language, script, and tokenization.

- Fine-Tuning: Chosen pre-trained models are fine-tuned using dataset provided by the organizers to develop primary systems with the addition of new vocabulary pertaining to these three languages for better representation.

- Data Augmentation and Contrastive systems development: Selectively incorporated related-language corpora provided by the organizer for primary submission. The contrastive submissions include supplementary corpora sourced from the web, generated synthetically (with and without utilizing monolingual corpora), and also drawn from proprietary data.

As part of synthetic parallel corpora generation back translation is carried out using openly available third party translation tool. ANVITA also employed pre-processing with set of language agnostic heuristics and selective post edits using LLM. On the WMT 2025 official test set, ANVITA achieved BLEU score of 2.41-11.59 with 2.2K to 60K corpora and 6.99-19.43 BLEU scores with the augmented corpora. Overall ANVITA ranked first for {Nyishi, Kokborok} ↔ English and second for Khasi ↔ English on the Official test set across all the evaluation metrics used by the organizer.

The rest of the paper is organized as given below. WMT 2025 task set up is described in 2. Section 3 presents Related work. Brief introduction to datasets is given in Section 4. ANVITA Indic LR System is described in Section 5. Section 5.5 describes the Experimental setup. Section 6 reports Evaluation results. Section 7 concludes the paper along with future directions.

2 Task Setup

WMT 2025 shared task on Low-Resource Indic Language comprised of translation of seven diverse, low resource Indian languages with the objective of developing robust MT systems that produce high-quality translations despite the constraints of data availability. The languages are divided into two categories. Category-1 comprised of five languages and 10 translation directions Assamese↔English, Mizo↔English, Khasi↔English, Manipuri↔English and Nyishi↔English with moderate training data and category-2 comprised of two languages, 4 directions Bodo↔English and Kokborok↔English with very limited training data. For each language pair three submissions were allowed. Primary systems with the constraint of using only the official data with additional monolingual resources and public pretrained models and two optional unrestricted Contrastive systems which may use external or additional parallel corpora beyond the organizer provided corpora.

ANVITA team submissions included three languages (Khasi, Nyishi, Kokborok), six translation directions (Khasi↔English, Nyishi↔English,

Kokborok↔English) and 17 systems comprising six Primary and eleven Contrastive systems.

3 Related Work

Building quality machine translation systems for low resource languages remains a critical research area. NLLB (No Language Left Behind) (NLLB Team et al., 2022) is one such research initiative to address limitations of MT for low resource languages which built large-scale multilingual model for 200 languages (including few low-resource Indian languages) on curated multilingual corpora, using transformer-based architecture (Vaswani et al., 2017). NLLB showed strong zero-shot and few-shot translation capabilities. As, fine-tuning such large models remains computationally expensive, our approach involved use of Low-Rank Adaptation (LoRA) (Hu et al., 2022), a method for parameter-efficient fine-tuning of large models to significantly reduce the number of update parameters. NLLB uses sub-word tokenizers such as sentencepiece (Kudo and Richardson, 2018) or BPE. To have a token free approach towards wider inclusion, ByT5 (Xue et al., 2022) introduced a byte-level variant of the T5 model (Raffel et al., 2020), which operates at the byte level rather than the token or sub-word level. Our work builds on these foundations. We leverage the methodologies and insights gained from WMT 2024's Indic MT shared task (Pakray et al., 2024) efforts but extend them to the more challenging low resource languages like Kokborok and Khasi.

4 Dataset

The training dataset provided by the WMT 2025 organizer varies from 60,000 to as low as 2,269 sentences as summarized in Table-1. ANVITA Primary submissions used only parallel corpora provided by the WMT 2025 organizer. Contrastive systems however used additional parallel data curated as part of this work and is summarized in Table-2. For evaluation, the official test set used is summarized in Table-3

5 ANVITA Indic LR Machine Translation System

This section describes the design of ANVITA systems for three low resource languages Nyishi, Khasi and Kokborok.

5.1 Data Preprocessing

As part of data preprocessing, selected noise filtering as described in (Vegi et al., 2021, 2022) are applied on all the datasets as mentioned in Table 1 and 2 with the objective of reducing corpora noise and improve translation quality. Additionally, non-linguistic artifacts such as timestamps, nan/null etc. present in any language are also removed from the bi-text corpora. Finally, text is processed using moses decoder [3] for punctuation normalization of English Text, and unicode normalization on both Indic and English text. Statistics of corpora before and after preprocessing is captured in Table-4.

5.2 Data for Contrastive Systems

Our supplementary data collection strategy for Contrastive systems include harvesting of monolingual and parallel corpora from web. Further to create synthetic parallel corpora for augmentation, the harvested monolingual resources are back translated (Kudugunta et al., 2019) using openly available third party translation tool, where only sentences with up to 20 words length are considered. Additional data is curated for Khasi and Kokborok languages, as detailed in the Table-2.

5.3 Nyishi ↔ English Translation Systems

For Nyishi ↔ English primary systems, ByT5 pretrained language model (Xue et al., 2022) is fine-tuned with the organizer data. ByT5 does not natively support Nyishi language. Unlike typical T5 variants that use subword tokenization like SentencePiece or BPE, ByT5 operates directly at the byte level. For Contrastive systems, data synthesis technique is used to generate additional parallel corpora from the organizer provided data using a custom method, as described in the subsequent subsection.

5.3.1 Byte-level Tokenization

Byte level (Wei et al., 2021) language and script agnostic tokenizer is used for both Nyishi and English text inline with the ByT5 language model. This approach does not require vocabulary building. So no new vocabulary is added. ByT5 is explored for better adaptation and potential noise robustness.

[3] https://github.com/hplt-project/sacremoses

Language Pairs	Language Family	Language Script	Parallel Corpus (P)	Number of words	Number of unique words	Average sentence length (num_words)
Nyishi (njz) - English (en)	Sino-Tibetan	Latin	60,000	(njz) 324,105 (en) 338,278	(njz) 39,074 (en) 13,647	(njz) 5.4 (en) 5.6
Khasi (kha) - English (en)	Austroasiatic	Latin	26,000	(kha) 966,353 (en) 798,291	(kha) 8,123 (en) 12,778	(kha) 37.16 (en) 30.7
Bodo (brx) - English (en)	Sino-Tibetan	Devanagari	15,215	(bodo) 20,4947 (en) 227,408	(bodo) 32,039 (en) 30,546	(bodo) 13.47 (en) 14.94
Kokborok (trp) - English (en)	Sino-Tibetan	Latin	2,269	(trp) 51,261 (en) 55,498	(trp) 6,619 (en) 6,175	(trp) 22.59 (en) 24.45

Table 1: Statistics WMT 2025 parallel corpora provided by organizer for Nyishi, Khasi, Bodo, Kokborok and used for Primary submission

Language Pairs	Number of sentences	Description	System: Contrastive 1	System: Contrastive 2
Nyishi (njz)-English (en)	30000	1. Curated parallel sentences by employing custom data synthesis technique from Nyishi-English parallel corpus provided by the WMT 2025 organizer with pairwise cosine similarity ≥ 0.8	Yes	No
	10000	2. Curated parallel sentences by employing custom data synthesis technique from Nyishi-English parallel corpus provided by the WMT 2025 organizer with pairwise cosine similarity ≥ 0.9	No	Yes
Khasi (kha)-English (en)	10,000	1. English sentences harvested from children stories and back translated to Khasi.	Yes	Yes
	5,50,000	2. Khasi sentences harvested from news portals and translated to English using back translation technique	Yes	No
	50,000	3. English sentences (taken from Nyishi-English parallel corpus provided by the WMT 2025 organizer) and back translated to Khasi	Yes	Yes
	7000	4. Sentences drawn from proprietary parallel corpora	Yes	Yes
Kokborok (trp)-English (en)	22793	1. Parallel sentences harvested from web	No	Yes
	50000	2. English sentences (taken from Nyishi-English parallel corpus provided by the WMT 2025 organizer) back translated to Kokborok	Yes	No

Table 2: Statistics of supplementary parallel corpora used for contrastive systems

5.3.2 Data Augmentation through Synthesis

In the organizer provided Nyishi-English parallel corpus, average length of sentences is 5.4 and 5.6 for Nyishi and English text respectively. So to create synthetic data involving long sentences from the existing data, a custom data synthesis technique is implemented as described below. The goal is to generate longer and coherent sentence pairs by concatenating compatible bilingual text segments, thereby enriching the training dataset.

Given a bilingual corpus of sentence pairs (S_i, T_i), where S_i is a source sentence and T_i is the corresponding target translation, we combine two such pairs (S_i, T_i) and (S_j, T_j) to form:

$$S_{ij} = S_i + joiner + S_j,$$

$$T_{ij} = T_i + joiner + T_j,$$

only if for S_i and S_j sentence-level cosine similarity is $\geq$ Threshold

- For Contrastive-1 system, *threshold* is chosen to be ≥ 0.8 , which resulted in 30000 parallel sentences. These synthetically created parallel sentences are augmented with the given training set.

- For Contrastive-2 system, *threshold* is chosen to be ≥ 0.9, which resulted into 10000 parallel sentences. These are augmented with the given training set.

Here *joiner* is usually sentence end marker followed by one white space.

The performance results of contrastive-1, contrastive 2 systems are presented in Table-7. Primary systems BLEU scores are better than the contrastive systems in case Nyishi-English language pair, indicating little effectiveness of the data synthesis method on the official test set.

Translation direction	Number of sentences in test file	Number of words	Number of unique words	Average sentence length (num_words)
English (en)→Nyishi (njz)	1,000 (en)	11,355	4,244	11.35
Nyishi (njz)→English (en)	1,000 (njz)	13,323	3,347	13.32
English (en)→Khasi (kha)	1,000 (en)	11,355	4,244	11.35
Khasi (kha)→English (en)	1,000 (kha)	22,399	2,186	22.39
English (en)→Kokborok (trp)	1,000 (en)	11,355	4,244	11.35
Kokborok (trp)→English (en)	1,000 (trp)	13,675	3,563	13.67

Table 3: Statistics of WMT 2025 Official Test set for Khasi, Kokborok, Nyishi

Language Pairs	Parallel data (Pb)- Before data processing	Parallel data (Pa)- After data processing
Nyishi (njz)-English (en)	60,000	51,000
Khasi (kha)-English (en)	26,000	25,995
Bodo (bodo) - English (en)	15,215	12,765
Kokborok (trp)-English (en)	2,269	2,266

Table 4: Statistics of preprocessed parallel corpora

Optimizer	AdamW
learning rate	1e-5
learning rate scheduler	linear with warmup
precision	fp16
patience	5
maximum number of epochs	20
metric_for_best_model	CHRF++

Table 5: Training parameters

peft type	LORA
rank	64
lora alpha	128
lora dropout	0.1
target modules	all

Table 6: LoRA configuration

5.4 {Khasi, Kokborok} ↔ English Translation Systems

Our approach for {Khasi, Kokborok} ↔ English Primary systems involve finetuning of pretrained translation model NLLB-200-distilled-600M (NLLB Team et al., 2022) [4] with the organizer provided data. Language specific additional vocabulary for Khasi and Kokborok are also added to the NLLB vocabulary. For contrastive systems similar techniques are employed, but with additional parallel corpora as described in Table-2.

5.4.1 Data Augmentation Through Related Language Data

For augmenting Kokborok →English training data for primary submission, Bodo↔English corpora provided by the organizer is utilized, as Bodo is related to Kokborok. For better transfer, Bodo text is converted from Devanagri script to Latin script using a romanization tool[5].

5.4.2 Data Augmentation Through Harvesting from Web and Synthesis

For contrastive systems, quality monolingual corpora is compiled from children stories, news portals and also used Nyshi↔English data provided by the organizer. Further, these sentences are back translated using openly available 3rd party translation tool. Sentences are also drawn from proprietary corpora as described in Table-2. As low resource languages Khasi and Kokborok do not have sentence tokenizers, pySBD (Sadvilkar and Neumann, 2020) is used for the same.

5.4.3 Sub-word Tokenization

Sub-word tokenization is one of critical component of any NMT system. In our work, specifically for building Khasi and Kokborok MT systems, SentencePiece (Unigram language model) tokenizer is trained on data provided by the organizer. The SentencePiece (Unigram language model) is chosen as it is compatible with NLLB tokenizer.

[4]https://github.com/facebookresearch/fairseq/tree/nllb

[5]https://github.com/anoopkunchukuttan/indic_nlp_library

Lang. Direction	Primary/ Contrastive	BLEU	METEOR	ROUGE-L	chrF	TER	Cos Similarity	Rank
English (en)→Nyishi (njz)	primary	6.21	0.21	0.28	34.01	81.53	-	First
	contrastive	5.92	0.20	0.27	34.08	82.83	-	First
Nyishi (njz)→English (en)	primary	11.59	0.41	0.51	49.85	74.09	0.79	First
	contrastive-1	11.13	0.42	0.51	48.92	74.17	0.80	First
	contrastive-2	11.25	0.40	0.51	49.36	73.79	0.78	-
English (en)→Khasi (kha)	primary	7.34	0.25	0.34	28.34	75.77	-	Second
	contrastive-1	18.83	0.45	0.54	45.48	55.75	-	-
	contrastive-2	19.43	0.46	0.55	45.93	54.41	-	Third
Khasi (kha)→English (en)	primary	1.99	0.11	0.14	20.88	223.26	0.30	Second
	contrastive-1	7.44	0.38	0.42	41.85	102.87	0.74	Second
	contrastive-2	4.39	0.22	0.28	30.65	123.25	0.55	-
English (en)→Kokborok (trp)	primary	1.76	0.11	0.17	18.58	104.04	-	First
	contrastive-1	6.99	0.30	0.37	38.08	76.26	-	First
	contrastive-2	0.55	0.04	0.05	13.38	335.55	-	-
Kokborok (trp)→English (en)	primary	2.41	0.11	0.18	23.55	129.15	0.36	First
	contrastive-1	2.99	0.16	0.22	25.52	117.73	0.49	First
	contrastive-2	0.79	0.05	0.08	16.46	170.6	0.20	-

Table 7: Performance of ANVITA on the WMT 2025 Official Test set

5.4.4 Fusion of Language Vocabulary with Pre-trained Model

To have better representation of language and reduce OOV words, NLLB vocabulary is updated with the sub-word vocabulary of Khasi and Kokborok languages. Initial vocabulary of NLLB tokenizer is 2,56,204, which is augmented with additional 831 Khasi vocabulary and 235 kokborok vocabulary and this took the final tally of NLLB vocabulary to 2,57,272. As NLLB does not support Khasi and Kokborok, hence special language tokens are also added to NLLB. This vocabulary inclusion enabled effective finetuning and better representation of words from low-resource languages.

5.4.5 {Khasi, Kokborok} ↔ English Model Training

The training steps followed for the four directions are as given below:

- Addition of Khasi and Kokborok language codes to NLLB vocabulary.
- Addition of Khasi and Kokborok sub-word vocabulary to NLLB vocabulary.
- For Kokborok primary systems, Bodo text is Romanized and augmented as presented in Table 1 .
- Primary models are trained on the data as presented in Table 1.
- Supplementary training data of contrastive systems include the data collected and back-translated as presented in Table 2.
- For Kokborok→English both contrastive models, Kokborok language token embedding is shared with the pre-trained NLLB Mizo (lus) language token embedding. Thus, the learned embedding is used for Kokborok token generation during inference time.
- Each language direction is separately fine-tuned on NLLB-200-distilled-600M (NLLB Team et al., 2022) model using Low-Rank Adaptation (LoRA) (Hu et al., 2022) method.
- For {Kokborok, Khasi} → English translation, English language token is set as target language and also it is forced as beginning of sentence. Similarly for other direction, corresponding language tokens are forced as beginning of sentence tokens. This is required since NLLB being a multilingual many to many MT model one needs to ensure tokens from the correct target language are generated.
- For Kokborok→English translation in Contrastive-2 submission, DeepSeek-R1 LLM (DeepSeek-AI et al., 2025) is used for post-editing of English translations.

5.5 Model Training and Experiment Details

All the experiments are conducted on NVIDIA DGX machine with 8xA100 80GB GPU cards.

ANVITA Indic LR used huggingface [6] toolkits for training. Training parameters and LORA configuration for all the experiments are shown in Table-5 and Table-6 respectively. For {Nyishi, Khasi} ↔ English, training batch-size is set to 32, whereas for Kokborok ↔ English training batch-size 16 is used. Gradient accumulation is used to avoid out-of-memory issues while training.

6 Evaluation and Result Analysis

Performance evaluation was carried out by the WMT 2025 organizer using BLEU, METEOR, ROUGE-L, chrF, TER and Cosine Similarity metrics on the Official test set (Table-3) for both Primary and Contrastive systems submitted. The results published by the organizers are shown in the Table-7.

Primary Systems: Nyishi→English with 60K parallel corpora attained BLEU score of 11.59, English→Khasi with 26K parallel corpora attained BLEU score of 7.34 and Kokborok→English with only 2.2K parallel corpora attained BLEU score of 2.41.

Contrastive Systems: With augmented corpora, overall English→Khasi achieved BLEU score of 19.43 and English→Kokborok 6.99.

Scores of our system for {Nyishi, Kokborok}→English directions are relatively better than that of English→{Nyishi, Kokborok} direction; English→ Khasi performance scores are better than Khasi→English direction.

In terms of ranks, ANVITA on the WMT 2025 official test set, secured First rank for {Nyishi, Kokborok} ↔ English and second for Khasi ↔ English across evaluation metrics including, BLEU, METEOR, ROUGE-L, chrF, TER and Cosine Similarity.

7 Conclusion

WMT 2025 shared task on Low-Resource Indic Language Translation posed significant challenging problem of building robust MT system for extremely low-resource Indian languages, many of which have little presence in digital domain. ANVITA team utilizing multi-pronged approach with transfer learning strategy achieved BLEU score of 2.41-11.59 for 2.2K to 60K corpora and secured top ranks. Overall with augmented data, the team achieved BLEU score of 6.99-19.43 which also fell short of robust MT system threshold indicating need for data augmentation in terms of both quality corpora and synthetic data and innovative techniques for data generation, suitable architectural enhancement and better learning objectives.

[6] https://huggingface.co/

Acknowledgments

The authors would like to thank Director, CAIR for his constant encouragement and supports. The authors would also like to thank Prasanna Kumar KR and Chitra Viswanathan for their guidance and enablement.

References

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian

Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nipun Sadvilkar and Mark Neumann. 2020. PySBD: Pragmatic sentence boundary disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Prasanna K R, and Chitra Viswanathan. 2022. ANVITA-African: A multilingual neural machine translation system for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1090–1097, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Chitra Viswanathan, and Prasanna Kumar K R. 2021. ANVITA machine translation system for WAT 2021 MultiIndicMT shared task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 244–249, Online. Association for Computational Linguistics.

Junqiu Wei, Qun Liu, Yinpeng Guo, and Xin Jiang. 2021. Training multilingual pre-trained language model with byte-level subwords.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

DoDS-IITPKD:Submissions to the WMT25 Low-Resource Indic Language Translation Task

Ontiwell Khongthaw* **G.L. John Salvin** **Shrikant Tryambak Budde**

Abigairl Nyasha Chigwededza **Dhruvadeep Malkar** **Swapnil Hingmire**

Mehta Family School of Data Science and Artificial Intelligence,
Department of Data Science, Indian Institute of Technology (IIT) Palakkad, Kerala, India
{142503002,142402010,142402015,142201026}@smail.iitpkd.ac.in,
okhongthaw@gmail.com, swapnilh@iitpkd.ac.in

Abstract

Low-resource translation for Indic languages poses significant challenges due to limited parallel corpora and linguistic diversity. In this work, we describe our participation in the WMT25 shared task for four Indic languages-Khasi, Mizo, Assamese, which is categorized into Category 1 and Bodo in Category 2. For our PRIMARY submission, we fine-tuned the distilled NLLB-200(600M) model on bidirectional English↔Khasi and English↔Mizo data, and employed the IndicTrans2 model family for Assamese and Bodo translation. Our CONTRASTIVE submission augments training with external corpora from PMINDIA, Google SMOL and GATITOS to further enrich low-resource data coverage. Both systems leverage Low-Rank Adaptation (LoRA) within a parameter-efficient fine-tuning framework, enabling lightweight adapter training atop frozen pretrained weights. The translation pipeline was developed using the Hugging Face Transformers and PEFT libraries, augmented with bespoke preprocessing modules that append both language and domain identifiers to each instance. We evaluated our approach on parallel corpora spanning multiple domains: article based, newswire, scientific, and biblical texts as provided by the WMT25 dataset, under conditions of severe data scarcity. Fine-tuning lightweight LoRA adapters on targeted parallel corpora yields marked improvements in evaluation metrics, confirming their effectiveness for cross-domain adaptation in low-resource Indic languages.

1 Introduction

Low-resource language translation remains one of the most persistent challenges in machine translation (MT), particularly for linguistically diverse regions such as India. We observed that in WMT25 the provided corpora spanned biblical, scientific, news, and article-based domain, introducing significant domain shifts that demanded robust adaptation strategies (Pakray et al., 2025). To address these challenges, we developed two primary systems. The first leveraged IndicTrans2, a transformer-based multilingual model optimized for Indic languages, and the second utilized NLLB-200(600M), a distilled multilingual model trained on over 200 languages. Both systems were fine-tuned using Low-Rank Adaptation (LoRA), enabling efficient domain adaptation without retraining the full model. For our contrastive submission, we augmented the training data with external corpora from sources such as PMINDIA (Haddow and Kirefu, 2020), GATITOS (Jones et al., 2023), and Google SMOL (Caswell et al., 2025), allowing us to explore the impact of data diversity on translation quality. This paper presents our system architecture, training methodology, and evaluation results, with a particular focus on how domain-specific corpora and external augmentation influence performance across four low-resource Indic languages: Khasi, Mizo, Assamese, and Bodo. Our approach employs parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA) on a pre-trained MT model, enabling a detailed empirical analysis of how large-scale architectures can be effectively adapted for low-resource languages under severe data constraints. The findings contribute to the growing body of research on scalable and adaptable MT systems for underrepresented languages.

2 Related Work

Translation quality in low-resource scenarios has been significantly advanced by large-scale multilingual models and lexical augmentation techniques. Fan et al. (2022) introduced No Language Left Behind (NLLB) which demonstrates effective multilingual MT at scale using a Sparsely Gated Mixture of Expert models trained with data that is mined specifically for underrepresented languages.

*Work done at IIT Palakkad.

Proceedings of the Tenth Conference on Machine Translation, pages 1248–1252
November 8-9, 2025

Their approach achieved substantial BLEU improvements and incorporated safety evaluations using FLORES-200 (Fan et al., 2022). Also, Jones et al. (2023) explored bilingual lexica as a lightweight data augmentation method, showing that collected lexical resources such as GATITOS can significantly enhance performance in unsupervised translation settings.

Toolkits like the HuggingFace Datasets library (Lhoest et al., 2021) also made efforts to support data development and reproducibility, which standardizes access to hundreds of multilingual corpora used in MT research.

For evaluation, several automatic metrics have been proposed to correlate better with human judgments. Lin (2004) developed ROUGE, widely used in summarization but also adopted in MT, which computes n-gram overlap and has influenced newer evaluation benchmarks. Banerjee and Lavie (2005) introduced METEOR, which matches unigrams using surface forms, stems and synonyms, incorporating both precision and recall as well as word order. Snover et al. (2006) proposed Translation Edit Rate (TER), also called Translation Error Rate, which measures the number of edits required to change a system output into one of the references. Popović (2015) proposed chrF, a character n-gram F-score metric that outperforms word-level metrics in many segment-level evaluations.

3 Dataset

For our primary submission, we utilized the Indic Machine Translation corpus from the WMT25 Shared Task. This benchmark comprises parallel data for four low-resource Indian languages, stratified into two categories based on training data volume. Category 1 encompasses language pairs with moderate-sized corpora, whereas Category 2 contains the severely data-starved corpora.

The language pairs are delineated as follows:

Category 1: en-as (English ↔ Assamese), en-lus (English ↔ Mizo), en-kha (English ↔ Khasi)

Category 2: en-bodo (English ↔ Bodo)

The parallel corpora supplied by the WMT25 IndicMT shared task[1] were employed for all model development. Each language pair's dataset was randomly divided into training (70 %), validation (20 %), and internal test (10 %) subsets, as detailed in Table 1. In addition, the task organizers released held-out monolingual test sets containing 1,000 sentences per translation direction for each language pair; these sets were used exclusively for final evaluation.

Language	Total Sentences	Train (70%)	Valid (20%)	Test (10%)
Assamese	54,000	37,800	10,800	5,400
Khasi	26,000	18,200	5,200	2,600
Mizo	50,000	35,000	10,000	5,000
Bodo	15,215	10,651	3,043	1,521

Table 1: Summary of Parallel Training Data from the WMT25 Indic MT Dataset.

3.1 Contrastive System Dataset

For a comparative analysis of data augmentation, we constructed a contrastive system by supplementing the WMT25 training dataset with additional publicly available parallel corpora. Our goal was to assess the resulting impact on translation performance across low-resource language pairs.

We incorporated data from four primary sources: the PMINDIA corpus (Haddow and Kirefu, 2020), high-quality parallel corpora for multiple Indian languages, sourced from government websites, official publications, and other public domain materials, covering legal, administrative, and general-purpose domains.; the GATITOS dataset (Jones et al., 2023), which provides lexically-augmented data for multilingual translation; the SMOL dataset (Caswell et al., 2025), containing professionally translated sentences for under-represented languages; and the Tatoeba corpus (Tiedemann, 2020), a large, community-sourced collection of multilingual sentence pairs.

The total volume of parallel data for each language after augmentation is detailed in Table 2. This table delineates the contribution of each external corpus alongside the original WMT data.

Corpus	Assamese (asm)	Bodo (brx)	Khasi (kha)	Mizo (lus)
WMT	54,000	15,216	26,000	50,000
GATITOS	3,975	3,994	4,000	3,998
Smol Sent	0	863	0	863
PMINDIA	9,732	0	0	0
Tatoeba	0	0	1,426	0
Total	67,707	20,073	31,426	54,861

Table 2: Parallel Corpus Statistics for the Contrastive System, detailing the original WMT25 data and supplementary corpora.

[1] https://www2.statmt.org/wmt25/indic-mt-task.html

4 Methodology

Our methodology is focused on fine-tuning state-of-the-art, pre-trained multilingual translation models that excel in low-resource settings. We chose NLLB-200(600M) (Fan et al., 2022) and IndicTrans2 (Gala et al., 2023) as our core architectures. NLLB-200(600M) , developed under the No Language Left Behind initiative, delivers extensive typological coverage and consistently high translation quality across diverse languages (Fan et al., 2022). IndicTrans2, by contrast, incorporates script-aware tokenization and subword segmentation tailored specifically to Indian languages, yielding superior performance on Indic↔English pairs (Gala et al., 2023).

By fine-tuning these complementary models on the WMT25 IndicMT parallel corpora and on the augmented corpus for our contrastive system, we established a strong performance baseline and systematically quantified the gains afforded by data augmentation.

4.1 Preprocessing

We employed a three-step preprocessing pipeline to ensure data consistency and compatibility with our models:

1. **Text Normalization:** English segments were processed using the `MosesPunctNormalizer` (Koehn et al., 2007), while a custom function (`preproc()`) performed Unicode NFKC normalization and non-printable character removal for Khasi and Mizo.

2. **Language Tagging:** Each sentence was prepended with a language-specific tag (e.g., `<eng_Latn>`, `<kha_Latn>`) to guide the multilingual model during fine-tuning.

3. **Dataset Structuring:** The processed sentence pairs were structured into a Hugging Face `DatasetDict` (Lhoest et al., 2021), enabling efficient batching, shuffling, and training via the `Trainer` API (Wolf et al., 2020).

4.2 System Description

4.2.1 Primary Submission

Our primary systems are based on fine-tuning two state-of-the-art multilingual models—NLLB-200(600M) and IndicTrans2—selected for their complementary strengths on low-resource and Indic-script translations.

NLLB-200(600M) for Khasi and Mizo: We adopted the `facebook/nllb-200-distilled-600M` checkpoint (Fan et al., 2022) for Khasi and Mizo tasks.

Model & Tokenizer: The standard `NLLBTokenizer` handles Mizo without modification; for Khasi we registered a new language token (`<kha_Latn>`) at token ID 256204 to correctly signal the source and target language.

LoRA Fine-Tuning: We applied Low-Rank Adaptation (LoRA) to all linear layers, updating only adapter weights. This approach enables efficient domain adaptation with fewer trainable parameters compared to full fine-tuning. Training ran for 30 epochs under Adafactor (learning rate 1×10^{-5}, batch size 32) with early stopping after 10 evaluations. Evaluation metrics were BLEU, METEOR, ROUGE-L, chrF and TER. Detailed LoRA hyperparameters appear in Table 4.

IndicTrans2 for Bodo and Assamese: For Bodo and Assamese, we used the `ai4bharat/indictrans2-indic-en-dist-200M` model (Gala et al., 2023), which employs an `IndicProcessor` to prepend language tokens such as `<brx_Deva>` and `<asm_Beng>`.

LoRA Fine-Tuning: We mirrored the NLLB-200(600M) setup (Adafactor, 1×10^{-5} learning rate, 32-sentence batch, 30 epochs, early stopping) and applied identical LoRA settings (see Table 4). The resulting adapter checkpoints are saved as lightweight artifacts.

4.2.2 Contrastive Submission

To quantify the effect of data augmentation, we retrained the same base models on extended parallel corpora. The tokenization and training pipeline remained identical, with two key LoRA adjustments to accommodate the increased data volume.

Model Setup: We reused NLLB-200(600M) for Khasi/Mizo (Fan et al., 2022) and IndicTrans2 for Bodo/Assamese (Gala et al., 2023). All supplementary bilingual data underwent the preprocessing and language-tagging workflow described in Section 3.

LoRA Adaptation: We increased the LoRA rank to 64 and α to 128 to provide greater adaptation capacity for the contrastive data, while retaining LoRA's parameter efficiency. Training was reduced to 15 epochs (Adafactor, 1×10^{-5} learning rate,

Direction	BLEU		METEOR		ROUGE-L		chrF		TER	
	P	C	P	C	P	C	P	C	P	C
as-en	21.40	21.75	0.695	0.690	0.701	0.703	66.14	65.77	54.90	53.77
en-as	17.54	17.64	0.422	0.422	0.007	0.007	57.75	57.71	71.17	74.81
kha-en	4.31	5.52	0.239	0.289	0.293	0.349	31.33	34.85	131.86	113.30
en-kha	14.20	20.08	0.370	0.452	0.431	0.534	39.95	47.36	87.50	59.98
lus-en	10.38	11.81	0.537	0.544	0.576	0.581	55.09	55.17	86.84	74.39
en-lus	14.26	14.72	0.415	0.407	0.515	0.506	48.51	48.55	72.22	69.49
bodo-en	21.68	22.11	0.627	0.629	0.679	0.688	62.95	63.55	54.29	52.84
en-bodo	24.45	24.97	0.513	0.519	0.168	0.169	67.71	67.81	51.84	51.50

Table 3: Results for all language pairs: Primary Submission Results (P) vs Contrastive Submission Results (C).

batch size 32), as detailed in Table 4. Performance comparisons against the primary systems isolate gains attributable to data augmentation.

Parameter	Primary Submission	Contrastive Submission
Optimizer	Adafactor	
Learning rate	1×10^{-5}	
Epochs	30	15
Precision	bf16	
PEFT type	LoRA	
Rank (r)	16	64
Alpha (α)	32	128
Dropout	0.05	
Target modules	all linear layers	

Table 4: LoRA Configuration for Primary and Contrastive Submissions.

5 Results

We evaluate our system submissions on the WMT IndicMT shared task for four low-resource Indian languages: Assamese, Khasi, Mizo, and Bodo. Tables 3 presents the comprehensive results for our primary and contrastive submissions respectively across all bidirectional translation pairs. All systems are evaluated using standard automatic metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), chrF (Popović, 2015), and TER (Snover et al., 2006).

The results demonstrate that our contrastive submissions generally achieved better or comparable performance across most language pairs and metrics compared to the primary submissions.

6 Conclusion

In this paper, we described the DoDS-IITPKD submissions to the WMT25 Low-Resource Indic Language Translation Task. Our systems were designed for multiple Indic-English and English-Indic translation directions, focusing particularly on Category-I languages of NorthEast India. We explored a combination of pre-trained multilingual models (IndicTrans, NLLB-200(600M)),fine-tuning strategies and LoRA-based efficient adaptation. Future work will focus on more domain-robust adaptation and incorporating quality estimation for improved translation reliability.

7 Acknowledgments

We thank the organizers of the WMT25 Shared Task on Low-Resource Indic Language Translation for providing the datasets and evaluation framework. We also gratefully acknowledge the Department of Data Science, IIT Palakkad, for computing resources and infrastructure support.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, Djibrila Diane, and Solo Farabado Cissé. 2025. Smol: Professionally translated parallel data for 115 under-represented languages.

Angela Fan, Shruti Bhosale, Holger Schwenk, and et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jay Gala, Sahil Choudhary, Ajitesh Sharma, Vinay Nair, Anoop Kunchukuttan, Pratik Patel, Anirudh Srinivasan, and Anupam Singh. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume, Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Quentin Lhoest, Albert Villanova Villanova, Leandro von Werra, Victor Sanh, Lewis Debut, Julien Chaumond, Mariama Drame, Lewis Tunstall, Eduardo del Moral, Javier Soriano, et al. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on Low-resource Indic Languages Translation. In *Proceedings of the Tenth Conference on Machine Translation under Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Preliminary Exploration of Phrase-Based SMT and Multi-BPE Segmentations through Concatenated Tokenised Corpora for Low-Resource Indian Languages

Saumitra Yadav and **Manish Shrivastava**
MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India
saumitra.yadav@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper describes our methodology and findings in building Machine Translation (MT) systems for submission to the WMT 2025 Shared Task on Low-Resource Indic Language Translation. Our primary aim was to evaluate the effectiveness of a phrase-based Statistical Machine Translation (SMT) system combined with a less common subword segmentation strategy for languages with very limited parallel data. We applied multiple Byte Pair Encoding (BPE) merge operations to the parallel corpora and concatenated the outputs to improve vocabulary coverage. We built systems for the English–Nyishi, English–Khasi, and English–Assamese language pairs. Although the approach showed potential as a data augmentation method, its performance in BLEU scores was not competitive with other shared task systems. This paper outlines our system architecture, data processing pipeline, and evaluation results, and provides an analysis of the challenges, positioning our work as an exploratory benchmark for future research in this area.

1 Introduction

Machine Translation (MT) has advanced rapidly in recent years, primarily driven by neural architectures and the availability of large-scale parallel corpora. However, these benefits are often confined to high-resource languages (Koehn and Knowles, 2017), leaving many languages with little or no translation support (Gowda et al., 2021). The WMT 2025 Shared Task on Low-Resource Indic Languages Translation addresses this gap by focusing on languages spoken in India with scarce digital resources.

While many efforts adapt high-resource MT techniques to low-resource settings, direct transfer often fails due to the data-hungry nature of neural networks. This has led to research in areas such as word segmentation and other preprocessing strategies (Ding et al., 2019; Abid, 2020; Huck et al., 2017; Ortega et al., 2020; Lankford et al., 2021; Domingo et al., 2023; Lee et al., 2024) to make systems more viable under data constraints. Although our focus here is on bilingual MT, we acknowledge the rise of multilingual and decoder-only models. Our goal was to investigate how far statistical models, combined with multiple BPE segmentations (Poncelas et al., 2020), could be pushed in highly constrained settings.

Bilingual MT systems have been successfully developed for other under-represented languages, such as Cantonese–Mandarin (Liu, 2022), English–Luganda (Kimera et al., 2025), Wolof–French (Dione et al., 2022), Bavarian–German (Her and Kruschwitz, 2024), and English–Manipuri (Singh et al., 2023; Singh and Singh, 2022), often with transformer-based architectures and customised segmentation like BPE (Li et al., 2024).

Previous work (Yadav et al., 2019; Yadav and Shrivastava, 2021; Akhbardeh et al., 2021) has shown that, for some low-resource Indic languages, SMT can outperform NMT. For the WMT 2025 Shared Task (Pakray et al., 2025), we therefore chose SMT for our systems targeting English $\leftrightarrow$ {Assamese, Khasi, Manipuri}.

The organisers provided parallel corpora for English–Kokborok, English–Bodo, English–Nyishi, English–Manipuri, English–Khasi, English–Mizo, and English–Assamese, building on earlier iterations (Pakray et al., 2024). We set out to determine whether a robust, traditional method like SMT—enhanced with a multiple-BPE data augmentation technique—could remain a viable option in such low-resource scenarios. This paper describes our approach and analyses the performance of our submissions.

2 Background

Low-resource MT faces unique challenges due to the scarcity of high-quality parallel corpora. Data sparsity leads to out-of-vocabulary (OOV) issues

Proceedings of the Tenth Conference on Machine Translation, pages 1253–1258
November 8-9, 2025

and poor generalisation. While Neural Machine Translation (NMT) dominates for high-resource pairs, its high data requirements limit its applicability without substantial augmentation (Sennrich et al., 2016a) or multilingual transfer (Mahata et al., 2023; Johnson et al., 2017).

Phrase-based SMT is often more resilient to small data sizes. By learning from statistical alignments of phrase pairs, it can perform robustly with limited resources.

For languages in the Indic family, which often exhibit rich morphology, subword segmentation is an effective preprocessing step (Prabhugaonkar et al., 2014). BPE, in particular, balances word-level and character-level representations, reduces vocabulary size, and mitigates OOV problems. Inspired by Poncelas et al. (2020), we extended this idea by applying multiple BPE merge operations to produce diverse segmentations, concatenating them to create a richer training set.

3 Data

The shared task corpora were drawn from previous WMT datasets and new resources (Pal et al., 2023; Kakum et al., 2023; Pakray et al., 2024). After preprocessing, we obtained the training statistics shown in Table 1.

Language Pair	# Training Sentences
English–Khasi	26,000
English–Mizo	50,000
English–Assamese	54,000

Table 1: Training data statistics before data augmentation.

Our preprocessing steps included:

- For Latin-script languages: tokenisation, normalisation, and lowercasing using Moses (Koehn et al., 2007).
- For others: processing with the Indic NLP Library (Kunchukuttan, 2020).
- For each parallel corpus: training and applying BPE (Sennrich et al., 2016b) with merge operations of 500, 1000, 2000, and 3000.

The segmented corpora from each merge setting were concatenated and deduplicated (Poncelas et al., 2020), resulting in the statistics in Table 2.

Language Pair	# Training Sentences
English–Khasi	91,379
English–Mizo	186,918
English–Assamese	209,010

Table 2: Training data statistics after concatenation and deduplication of multi-BPE segmentations.

4 System Description

We used Moses (Koehn et al., 2007) for phrase-based SMT, with target-side KenLM language models (Heafield, 2011) trained on the corpora in Table 2. Each system was evaluated under four inference configurations:

- **1000 BPE** segmented source
- **2000 BPE** segmented source
- **3000 BPE** segmented source
- **Combined Hypothesis**, selecting the output with the highest probability among the above. We select translation that exhibit an average log-likelihood of -1.0 or higher, according to measurements taken by the fairseq-interactive tool.

5 Results

Our systems were tested on the official WMT 2025 English–Nyishi, English–Khasi, and English–Assamese sets. In all cases, performance ranked in the lower tier, with BLEU scores notably behind top-performing NMT systems that likely used external data or more advanced architectures. Full results are shown in Tables 3 and 4.

6 Discussion

The modest results highlight the limitations of SMT in this setting. Two main factors likely contributed:

1. SMT's inability to capture long-range dependencies and nuanced patterns compared to modern neural models.
2. The data augmentation via multiple BPE segmentations did not sufficiently overcome the extreme scarcity of parallel data, particularly for Nyishi.

Noisy and domain-specific terms in the provided corpora may have further impacted translation quality.

To English	Test Inferenceing Strategy	BLEU	METEOR	ROUGE-L	CHRF	TER	Cos Similarity
Assamese	Combine Hypothesis	0.3331	0.0229	0.0213	17.5032	286.2066	0.0775
	1000 BPE	0.3331	0.0229	0.0213	17.5032	286.2066	0.0775
	2000 BPE	0.3331	0.0229	0.0213	17.5032	286.2066	0.0775
	3000 BPE	0.3340	0.0230	0.0214	17.5031	286.2821	0.0775
Khasi	Combine Hypothesis	1.0536	0.0793	0.1112	19.4678	177.4341	0.2460
	1000 BPE	1.0935	0.0808	0.1138	19.2604	171.4254	0.2428
	2000 BPE	1.0604	0.0802	0.1111	19.4635	176.1309	0.2456
	3000 BPE	1.0461	0.0806	0.1114	19.5720	179.1631	0.2474
Nyishi	Combine Hypothesis	1.2657	0.0857	0.1209	23.4376	138.2275	0.2111
	1000 BPE	1.2557	0.0828	0.1191	23.2949	139.4495	0.2033
	2000 BPE	1.1942	0.0808	0.1158	22.9758	145.2652	0.2052
	3000 BPE	1.1885	0.0811	0.1127	23.3545	147.9239	0.2006

Table 3: Indic Language to English translation systems

English To	Test Inferenceing Strategy	BLEU	METEOR	ROUGE-L	CHRF	TER
Assamesse	Combine Hypothesis	2.9694	0.1126	0.0000	31.4566	107.3459
	1000 BPE	2.9256	0.1089	0.0000	30.4868	104.2561
	2000 BPE	3.0255	0.1137	0.0000	31.2960	107.3301
	3000 BPE	3.0287	0.1145	0.0000	31.6349	108.9135
Khasi	Combine Hypothesis	4.2570	0.1922	0.2555	26.7990	96.2399
	1000 BPE	4.2404	0.1885	0.2539	26.5474	94.7060
	2000 BPE	4.2277	0.1933	0.2550	26.7577	97.9426
	3000 BPE	4.0968	0.1938	0.2522	26.9003	100.6179
Nyishi	Combine Hypothesis	1.1870	0.0492	0.0781	20.3680	123.9258
	1000 BPE	1.2280	0.0493	0.0782	20.2094	120.4597
	2000 BPE	1.1843	0.0496	0.0771	20.4325	124.4046
	3000 BPE	1.1730	0.0503	0.0770	20.6541	127.1327

Table 4: English to Indian Language translation systems

7 Conclusion and Future Work

Our participation in the WMT 2025 Shared Task was an exploratory test of a multi-BPE augmentation strategy in an SMT framework for extremely low-resource Indic language pairs. While the method did not yield competitive results, it provides a clear baseline for SMT in these settings and reinforces the potential value of hybrid or neural approaches. Future work will explore SMT–NMT hybrids, fine-tuning large multilingual models on limited data, and advanced augmentation methods such as back-translation.

References

Wael Abid. 2020. The SADID evaluation datasets for low-resource spoken language machine translation of Arabic dialects. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6030–6043, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

Cheikh M. Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, and Sileye Ba. 2022. Low-resource neural machine translation: Benchmarking state-of-the-art transformer for Wolof<->French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6654–6661, Marseille, France. European Language Resources Association.

Miguel Domingo, Mercedes García-Martínez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2023. How much does tokenization affect neural machine translation? In *Computational Linguistics and Intelligent Text Processing*, pages 545–554, Cham. Springer Nature Switzerland.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Wan-hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using Bavarian as a case study. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 155–167, Torino, Italia. ELRA and ICCL.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017. Target-side word segmentation strategies for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 56–67, Copenhagen, Denmark. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nabam Kakum, Sahinur Rahman Laskar, Koj Sambyo, and Partha Pakray. 2023. Neural machine translation for limited resources english-nyishi pair. *Sādhanā*, 48(4):237.

Richard Kimera, DongNyeong Heo, Daniela N. Rim, and Heeyoul Choi. 2025. Data augmentation with back translation for low resource languages: A case of english and luganda. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 142–148, New York, NY, USA. Association for Computing Machinery.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Seamus Lankford, Haithem Alfi, and Andy Way. 2021. Transformers for low-resource languages: Is féidir linn! In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 48–60, Virtual. Association for Machine Translation in the Americas.

Jungseob Lee, Hyeonseok Moon, Seungjun Lee, Chanjun Park, Sugyeong Eo, Hyunwoong Ko, Jaehyung Seo, Seungyoon Lee, and Heuiseok Lim. 2024. Length-aware byte pair encoding for mitigating oversegmentation in Korean machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2287–2303, Bangkok, Thailand. Association for Computational Linguistics.

Fuxue Li, Beibei Liu, Hong Yan, Mingzhi Shao, Peijun Xie, Jiarui Li, and Chuncheng Chi. 2024. A bilingual templates data augmentation method for low-resource neural machine translation. In *Advanced Intelligent Computing Technology and Applications: 20th International Conference, ICIC 2024, Tianjin, China, August 5–8, 2024, Proceedings, Part III*, page 40–51, Berlin, Heidelberg. Springer-Verlag.

Evelyn Kai-Yan Liu. 2022. Low-resource neural machine translation: A case study of Cantonese. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 28–40, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Sainik Kumar Mahata, Dipanjan Saha, Dipankar Das, and Sivaji Bandyopadhyay. 2023. Transfer learning in low-resourced MT: An empirical study. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 646–650, Goa University, Goa, India. NLP Association of India (NLPAI).

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Partha Pakray, Reddi Krishna, Santanu Pal, Advaitha Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 shared task on low-resource indic languages translation. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025) under the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Suzhou, China.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Alberto Poncelas, Jan Buts, James Hadley, and Andy Way. 2020. Using multiple subwords to improve English-Esperanto automated literary translation quality. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 108–117, Suzhou, China. Association for Computational Linguistics.

Neha R Prabhugaonkar, Jyoti Pawar, Apurva S Nagvenkar, Pushpak Bhattacharyya, Diptesh Kanojia, and Manish Shrivastava. 2014. Panchbhoota: Hierarchical phrase based machine translation systems for five indian languages.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kshetrimayum Boynao Singh, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. NITS-CNLP low-resource neural machine translation systems of English-Manipuri language pair. In *Proceedings of the Eighth Conference on Machine Translation*, pages 967–971, Singapore. Association for Computational Linguistics.

Salam Michael Singh and Thoudam Doren Singh. 2022. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Syst. Appl.*, 209(C).

Saumitra Yadav, Vandan Mujadia, and Manish Shrivastava. 2019. A3-108 machine translation system for LoResMT 2019. In *Proceedings of the 2nd Workshop*

on Technologies for MT of Low Resource Languages, pages 64–67, Dublin, Ireland. European Association for Machine Translation.

Saumitra Yadav and Manish Shrivastava. 2021. A3-108 machine translation system for similar language translation shared task 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 304–306, Online. Association for Computational Linguistics.

AkibaNLP-TUT: Injecting Language-Specific Word-Level Noise for Low-Resource Language Translation

Shoki Hamada[1] **Tomoyoshi Akiba**[1] **Hajime Tsukada**[2]
[1]Toyohashi University of Technology [2] Aichi Sangyo University
{hamada.shoki.ew, akiba.tomoyoshi.tk}@tut.jp
tsukada@asu.ac.jp

Abstract

In this paper, we describes our system for the WMT 2025 Low-Resource Indic Language Translation Shared Task. The language directions addressed are Assamese↔English and Manipuri→English. We propose a method to improve translation performance from low-resource languages (LRLs) to English by injecting Language-specific word-level noise into the parallel corpus of a closely related high-resource language (HRL). In the proposed method, word replacements are performed based on edit distance, using vocabulary and frequency information extracted from an LRL monolingual corpus. Experiments conducted on Assamese and Manipuri show that, in the absence of LRL parallel data, the proposed method outperforms both the w/o noise setting and existing approaches. Furthermore, we confirmed that increasing the size of the monolingual corpus used for noise injection leads to improved translation performance.

1 Introduction

There are approximately 7,000 languages in the world, but only a small subset of high-resource languages (HRLs) have sufficiently developed parallel corpora for machine translation (MT). For these languages, research leveraging few-shot learning with parallel data (Zhu et al., 2023) and large-scale multilingual language models (mLLMs) (Xu et al., 2023; Zhou et al., 2023) has progressed. Such efforts have enabled the learning of shared cross-lingual embedding spaces, thereby facilitating cross-lingual transfer.

In contrast, many languages are low-resource languages (LRLs) for which only monolingual data is available. Compared to parallel data, monolingual data is easier to collect and is widely used for techniques such as back-translation (BT)(Sennrich et al., 2016a) and continued pretraining of mLLMs. This study aims to improve LRL→English translation accuracy by leveraging LRL monolingual data to inject language-specific, word-level noise into parallel data of closely related HRLs.

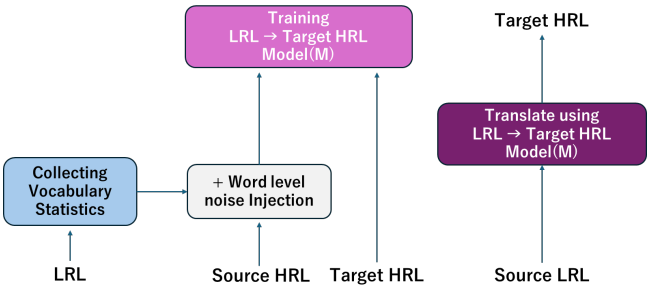


Figure 1: Overview of our proposed method: Language-Specific Word-Level Noise Injection

Maurya et al. (2024) proposed CharSpan, a method that injects character-level noise into HRL parallel data with high lexical similarity to a LRL. This method improves translation accuracy for the LRL by leveraging its character list. However, this approach does not sufficiently capture word-level statistical characteristics such as the LRL's vocabulary frequency and distribution.

To address this limitation, we propose a method that uses word lists and frequency information extracted from LRL monolingual data to add word-level noise to HRL parallel data. We further analyze the effect of this method under conditions where LRL parallel data is available versus unavailable. Experimental results show that when using only LRL monolingual data, our method outperforms existing approaches. In contrast, when LRL parallel data is available, the performance gap with existing methods is small. We also observe that increasing the size of the monolingual data used for noise injection tends to improve performance, and that including the test dataset in the monolingual data yields additional performance gains.

2 Related Work

Some studies have examined the effects of adding noise to parallel data on its diversity and the ro-

Proceedings of the Tenth Conference on Machine Translation, pages 1259–1264
November 8-9, 2025

bustness of translation models, but the impact on cross-lingual transfer has not been thoroughly investigated. Gal and Ghahramani (2016) proposed Word Dropout, which randomly sets some word embeddings to zero vectors. Wang et al. (2018) proposed a data augmentation method that adds random word replacements to parallel data. While both of these methods enhance the diversity of parallel data, their effectiveness in improving cross-lingual transfer capability is limited.

A representative data augmentation technique that leverages monolingual data in neural machine translation is back-translation (BT) (Sennrich et al., 2016a). When parallel data is scarce, BT generates pseudo-parallel data by using target-side monolingual data and a reverse-direction translation model. More recently, iterative back-translation (IBT) (Morita et al., 2018; Hoang et al., 2018; Zhang et al., 2018) has been proposed, which extends BT in both directions. IBT utilizes monolingual data from both sides to generate pseudo-parallel data in both directions and iteratively alternates between generating this data and updating the translation models in both directions.

3 Method

In this chapter, we propose a method to enhance robustness in LRL→En translation and promotes cross-lingual transfer by adding LRL-specific word-level noise into a parallel corpus of a closely related HRL. The noise consists solely of word replacements, where the edit distance is selected based on a geometric distribution. The replacement candidates are chosen using frequency-weighted selection, thereby injecting LRL words into the related HRL.

Specifically, we use approximately 160,000 sentences of Assamese monolingual text to add noise to Bengali–English parallel data. Figure 1 shows an overview of the proposed method. An example of the noise injection process is illustrated in Figure 2. Furthermore, by using the model trained on the noise-injected parallel data as a back-translation model, we perform En→LRL translation.

3.1 Language-Specific Word-level noise

We add word-level noise to the source-side training data D_{HRL} of the HRL pair to create the noisy parallel data D'_{HRL}.

First, we randomly select a word index x_i from any given sentence x. Next, we determine the edit distance d according to Equation 1, where p is the success probability of the geometric distribution:

HRL(Bn): এমন কথা কিন্তু তিনি আপনাকে কোনওদিনও বলেননি ।
Eng: But he never told me.

HRL(Bn) + Noise: এমন কথা সিল্কু তিনি আপনাকে কোনোদিনে বলেননি ।

Figure 2: Example of word-level noise injection for Bengali (HRL). The original Bengali sentence and its English sentence are shown at the top. In the noisy version (bottom), Bengali words are replaced with words selected from an Assamese vocabulary list.

$$P(d = k) \propto p(1-p)^{k-1} \quad (k = 1, 2, \ldots, K) \quad (1)$$

The parameters K and p control the distribution of noise magnitude. Then, based on the edit distance between a candidate word w and x_i, we extract a candidate set $V(d, x_i)$ from the LRL vocabulary V_{LRL}:

$$V(d, x_i) = \{w | w \in V_{LRL}, ED(w, x_i) = d\} \quad (2)$$

Here, $ED(\cdot, \cdot)$ denotes the Levenshtein distance. The replacement word w' is selected according to the product of $P(d)$ and the relative word frequency $f(w')$ in the LRL monolingual corpus:

$$P(w') = P(d) \cdot \frac{f(w')}{\sum_{w \in V(d, x_i)} f(w)} \quad (3)$$

This procedure is repeated until the proportion of characters changed by substitution in each sentence reaches a predefined target ratio.

3.2 Back translation

In this study, we apply translation model M, trained on HRL parallel data D'_{HRL} augumented with word-level noise, to LRL-to-English translation. Specifically, to perform EN→LRL translation, we first translate the LRL monolingual data into English using M, thereby creating pseudo-parallel data $\hat{D}_{\mathrm{LRL-EN}}$.

Subsequently, this pseudo-parallel data is used as input to train an English→LRL translation model.

4 Experimental Setup

4.1 Datasets

We target Bengali (Bn) as the HRL and Assamese (As) and Manipuri (Mni) as the LRLs, using multiple parallel and monolingual corpora for model

Corpora	Language	Usage	# Sentences
Samanantar	English-Bengali	Train	8,604,580
WMT25 Shared Task	English-Assamese	Train / Valid	Train: 53,003 Valid: 997
	English-Manipuri	Train / Valid	Train: 22,690 Valid: 997
FLORES-200	English-(Bengali, Assamese, Manipuri)	Valid / Test	Valid: 977 Test: 1,012
Community 2017 Wikipedia 2021	Assamese	Noise Injection	63,627 100,000

Table 1: Overview of the parallel and monolingual corpora used in this study, including the languages, their intended usage, and the number of sentences.

training and evaluation. Table 1 provides an overview of the corpora used. For parallel corpora, we used the large-scale English–Bengali Samanantar corpus(Ramesh et al., 2022) and the English–Assamese and English–Manipuri parallel datasets provided by the WMT25 shared task(Pal et al., 2023; Pakray et al., 2024). For evaluation, we used the validation and test sets of FLORES-200(Costa-Jussà et al., 2022). Additionally, Assamese Wikipedia 2021 and Community 2017 were used as monolingual corpora to extract vocabulary for noise injection in the proposed method.

4.2 Data Preprocessing

For English data, we first perform Unicode normalization (NFKC) and applied tokenization using `sacremoses`. Next, to reduce case variation at sentence beginnings and in proper nouns, we applied truecasing, and finally, we learned and applied Byte Pair Encoding (BPE)(Sennrich et al., 2016b; Gage, 1994) using `subword-nmt`. The number of BPE merge operations was set to 16,000.

For Bengali, Assamese, and Manipuri data, we applied the same preprocessing steps as for English—NFKC normalization, tokenization with `sacremoses`, and BPE (16,000 merges)—but did not apply truecasing.

4.3 Settings

For the word-level noise injection, the maximum edit distance K was set to 5, and the actual edit distance was sampled from a geometric distribution with a success probability $p = 0.5$. Noise was injected to each sentence until the proportion of characters altered by substitution reached the target ratio of 10%. Figure 4 shows the list of replacement characters used in CharSpan.

Parameter	Value
Architecture	Transformer (Encoder 6 layers / Decoder 6 layers)
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.98$)
Initial learning rate	$5e-4$
LR scheduler	Inverse Sqrt Decay
Gradient clip norm	1.0
Dropout	0.2
Max tokens / batch	8,000
Early-stopping patience	5 validations
GPU	2 × NVIDIA GeForce RTX 2080 Ti

Table 2: Model implementation and training details

We adopted the Transformer architecture (Vaswani et al., 2017) as the translation model. Both the encoder and decoder consisted of 6 layers, and optimization was performed using Adam (Kingma and Ba, 2014) with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The initial learning rate was set to 5×10^{-4}, with Inverse Sqrt Decay as the learning rate scheduler. To prevent gradient explosion, the gradient clip norm was set to 1.0. The dropout rate was set to 0.2, and the maximum number of tokens per batch was 8,000. Training employed early stopping, terminating when the validation loss did not improve for 5 consecutive evaluations. Experiments were conducted using two NVIDIA GeForce RTX 2080 Ti GPUs. Table 2 presents the key hyperparameter settings used in our experiments.

4.4 Evaluation Metrics

For validation and evaluation, we used the official FLORES-200 dev/test sets. BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017) were adopted as evaluation metrics.

5 Result and Analisys

5.1 Main Results: LRL→EN

Table 3 presents the translation results from LRL to English. **Bold** indicates the highest score in each setting. In the setting without LRL parallel

Models	As→En BLUE	As→En chrF	Mni→En BLUE	Mni→En chrF
w/o noise	5.49	25.2	1.29	**18.9**
CharSpan	10.44	36.1	**0.75**	18.3
Word-Level noise	**12.92**	**38.1**	0.65	17.3
w/o noise + parallel	19.07	44.4	9.8	34.8
CharSpan + parallel	**21.46**	**47.4**	11.97	**38.8**
Word-Level noise + parallel	21.44	46.9	**12.12**	37.3

Table 3: Experimental results of LRL→English translation with and without LRL parallel data.

# Sentences	Vocabulary Size
(full data + test) 164,639	182,727
(full data) 163,627	180,810
120,000	154,461
100,000	139,751
50,000	93,797
10,000	33,620

Table 4: Monolingual Assamese subsets used to build the LRL vocabulary for noise injection. Vocabulary size counts unique types after preprocessing.

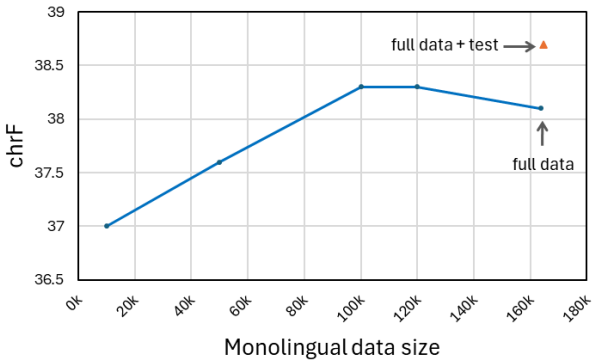


Figure 3: Effect of Assamese monolingual data size and domain on chrF scores for Assamese→English translation.

data, the proposed method outperformed both the w/o noise and CharSpan baselines for Assamese across all evaluation metrics. This improvement is likely due to the effective utilization of vocabulary and word frequency distributions derived from Assamese monolingual data. In contrast, for Manipuri, the proposed method underperformed compared to both baselines.

When LRL parallel data was used, the proposed method outperformed w/o noise for both languages, but achieved only comparable gains to CharSpan. This suggests that when w/o noise already possesses a moderate level of translation capability, the improvements brought by noise injection may be limited. Furthermore, for Manipuri, the presence or absence of LRL parallel data resulted in differing levels of improvement, indicating that the proposed method is effective when the w/o noise already has a certain degree of translation capability.

5.2 Effects of Monolingual Data Size and Domain for Noise Injection

The noise injection function used in this study (Equation 3) extracts replacement candidates from the LRL vocabulary with frequency weighting. Expanding the size of the monolingual corpus increases the likelihood of selecting more informative candidates. To verify this effect, we conducted experiments in which the amount of monolingual data was restricted. Starting from the full Assamese monolingual pool, we create down-sampled subsets with the following sentence counts (vocabulary sizes): *120k* (154,461), *100k* (139,751), *50k* (93,797) and *10k* (33,620). Additionally, we evaluate a setting where the full monolingual data is augmented with the Assamese test sentences from FLORES-200 (*full data + test*). For each subset, we learn BPE, train the model with the same configuration as described in Table 2, and evaluate it on FLORES-200.

The results are shown in the figure 3. The full data + test setting achieved the highest score of 38.7, slightly surpassing the 38.1 of the full data setting. The 120k and 100k settings both yielded similar performance at 38.3, while 50k achieved 37.6 and 10k scored 37.0, showing a gradual decline in performance as the amount of data decreased. These results suggest that increasing the size of the monolingual data makes it easier to select more informative replacement candidates during noise injection, potentially leading to improved translation performance. Furthermore, in the full data + test setting, including sentences from the same domain as the evaluation data in the monolingual corpus may have contributed to the performance improvement.

5.3 Shared Task Results

Table 5, presents the evaluation results of LRL↔English translation on the test set provided in the shared task. The evaluation metrics are BLEU, METEOR(Banerjee and Lavie, 2005), ROUGE-L(Lin, 2004), chrF, and TER(Snover et al., 2006). The translation directions are Assamese→English, English→Assamese, and Ma-

Direction	BLEU↑	METEOR↑	ROUGE-L↑	chrF↑	TER↓
Assamese→English	12.28	0.54	0.56	55.61	78.24
English→Assamese	14.03	0.38	0.01	53.76	74.08
Manipuri→English	5.74	0.33	0.37	41.28	109.95

Table 5: Evaluation results of LRL↔English translation using the test set provided in the shared task.

Language	Script	Candidate Alphabets
Assamese	Bengali	'ঁ', 'ং', 'ঃ', 'অ', 'আ', 'ই', 'ঈ', 'উ', 'ঊ', 'ঋ', 'এ', 'ঐ', 'ও', 'ঔ', 'ক', 'খ', 'গ', 'ঘ', 'ঙ', 'চ', 'ছ', 'জ', 'ঝ', 'ঞ', 'ট', 'ঠ', 'ড', 'ঢ', 'ণ', 'ত', 'থ', 'দ', 'ধ', 'ন', 'প', 'ফ', 'ব', 'ভ', 'ম', 'য', 'ল', 'শ', 'ষ', 'স', 'হ', '়', 'া', 'ি', 'ী', 'ু', 'ূ', 'ৃ', 'ে', 'ৈ', 'ো', 'ৌ', '্', 'ৎ', 'ড়', 'ঢ়', 'য়', '০', '১', '২', '৩', '৪', '৫', '৬', '৭', '৮', '৯', 'ৰ', 'ৱ'

Figure 4: Candidate alphabets for Assamese in the Bengali script, used as noise for CharSpan.

nipuri→English.

Our system was particularly designed for the Assamese→English direction. The performance gap between Assamese and Manipuri reflects the disparity in the availability of monolingual and parallel data, as well as differences in vocabulary and grammar with Bengali. These results indicate that the proposed method is especially effective for LRLs that are lexically close to the corresponding HRL.

6 Conclusion

In this study, we proposed Language-specific word-level noise injection method for the parallel corpus of a HRL closely related to a LRL, using vocabulary and frequency information extracted from the LRL's monolingual data. Experiments on Assamese and Manipuri demonstrated that, particularly in the absence of parallel data, the proposed method outperforms both the w/o noise setting and existing approaches. Furthermore, we showed that increasing the size of the monolingual data contributes to improved translation performance.

Limitations

First, the proposed method targets only a limited set of LRLs (Assamese and Manipuri), and its applicability to other languages has not been verified. In particular, it may be difficult to apply the method to LRLs whose script systems and lexical structures differ substantially from those of the HRL. Furthermore, the selection of replacement candidates for noise injection relies on word frequency and edit distance derived from the monolingual corpus of the LRL, and the potential for performance improvement may be limited depending on the scale of these statistics and the domain of the corpus.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kaushal Maurya, Rahul Kejriwal, Maunendra Desarkar, and Anoop Kunchukuttan. 2024. CharSpan: Utilizing lexical similarity to enable zero-shot machine translation for extremely low-resource languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 294–310, St. Julian's, Malta. Association for Computational Linguistics.

Tomohiro Morita, Tomoyosi Akiba, and Hajime Tsukada. 2018. A study on unsupervised adaptation of neural machine translation with bidirectional back-translation (in Japanese). In *IPSJ SIG Technical Report*, volume 2018-NL-238, pages 1–5.

Partha Pakray, Santanu Pal, Advaitha Vetagiri, Reddi Krishna, Arnab Kumar Maji, Sandeep Dash, Lenin Laitonjam, Lyngdoh Sarah, and Riyanka Manna. 2024. Findings of WMT 2024 shared task on low-resource Indic languages translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 654–668, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, Brussels, Belgium. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Joint training for neural machine translation models with monolingual data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

BVSLP: Machine Translation using Linguistic Embellishments for IndicMT Shared Task 2025

Nisheeth Joshi[1#], Palak Arora[2*], Anju Krishnia[1‡], Riya Lonchenpa[1], Mhasilenuo Vizo[1±]**
[1]Speech and Language Processing Lab, Banasthali Vidyapith, Rajasthan, India
[2]School of Computing, DIT University, Uttarakhand, India
[#]nisheeth.joshi@rediffmail.com, [*]palak.arora.pa55@gmail.com, [‡]aasvkrishnia@gmail.com,
[**]lonchenpar@gmail.com, mhasilenuovizo3@gmail.com

Abstract

This paper describes our submission to the Indic MT 2025 shared task, where we trained machine translation systems for five low-resource language pairs: English–Manipuri, Manipuri–English, English–Bodo, English–Assamese, and Assamese–English. To address the challenge of out-of-vocabulary errors, we introduced a Named Entity Translation module that automatically identified named entities and either translated or transliterated them into the target language. The augmented corpus produced by this module was used to fine-tune a Transformer-based neural machine translation system. Our approach, termed HEMANT (Highly Efficient Machine-Assisted Natural Translation), demonstrated consistent improvements, particularly in reducing named entity errors and improving fluency for Assamese–English and Manipuri–English. Official shared task evaluation results show that the system achieved competitive performance across all five language pairs, underscoring the effectiveness of linguistically informed preprocessing for low-resource Indic MT.

1 Introduction

This paper presents our submission to the IndicMT 2025 shared task, where we developed machine translation (MT) systems for five language pairs: English–Assamese, Assamese–English, English–Manipuri, Manipuri–English, and English–Bodo. The systems follow a two-stage pipeline comprising preprocessing and neural machine translation (NMT). In preprocessing, the source text undergoes tokenization, spelling normalization, and source-side linguistic analysis, with particular focus on identifying and translating named entities into the target language to reduce out-of-vocabulary (OOV) errors. The processed corpus is then used to train NMT models based on an encoder–decoder architecture, with subword segmentation applied via the SentencePiece tokenizer using the Byte Pair Encoding (BPE) model to improve vocabulary coverage and generalization. The resulting systems integrate named entity handling with standard NMT methods, thereby enhancing translation quality across the five language pairs.

2 Related Work

Research in Indic machine translation has increasingly emphasized the use of multilingual pretrained models and transfer learning to overcome data scarcity. The IndicTrans system (Ahuja et al., 2022) demonstrated the effectiveness of multilingual Transformer pretraining for Indian languages, providing strong baselines for Indo-Aryan and Dravidian pairs. More recently, No Language Left Behind (NLLB-200) (Costa-jussà et al., 2022) has scaled this paradigm further, offering pretrained models across 200 languages, including Assamese, thereby enabling robust fine-tuning for low-resource Indic settings.

The benefits of transfer learning for low-resource machine translation have been well-documented. Zoph and Knight (2016) showed that multi-source translation improves performance by leveraging related languages, while Nguyen and Chiang (2017) highlighted the effectiveness of cross-lingual transfer in low-resource neural MT. In the Indic context, transfer between closely related languages such as Bengali and Assamese or

Proceedings of the Tenth Conference on Machine Translation, pages 1265–1270
November 8-9, 2025

Manipuri and Bodo provides an opportunity to exploit linguistic similarities for improved translation performance.

Subword segmentation strategies have also been an important focus. Sennrich et al. (2016) introduced Byte Pair Encoding (BPE) as a means to mitigate out-of-vocabulary issues, while Kudo and Richardson (2018) proposed the Unigram Language Model as an alternative. More recently, Ahmed et al. (2023), in the WMT 2023 shared task, compared segmentation schemes across low-resource languages and showed that alternative approaches can outperform BPE in certain settings.

Efforts to improve named entity handling in MT are relatively fewer. Joshi and Katyayan (2023) and Sharma et al. (2023) demonstrated that augmenting training corpora with entity translations significantly reduces OOV errors in English–Braille and Hindi–English systems, respectively. Our work extends this line by systematically integrating a Named Entity Translation module into the preprocessing pipeline for Indic language pairs.

3 System Description

3.1 Data Preprocessing

Text tokenization and spelling correction were first performed on the source language corpus. Spelling normalization was applied to reduce orthographic inconsistencies in the corpora, particularly for Assamese and Manipuri. For Assamese, Unicode normalization was enforced, and common spelling variants arising from character duplication or visually similar graphemes were standardized using a manually crafted rule set. For Manipuri and Bodo, we implemented an edit-distance–based correction method (Levenshtein distance) supplemented with frequency statistics from the training corpus. Candidate corrections were chosen from a lexicon compiled from Wikipedia dumps, news portals, and government gazetteers, with the most frequent form selected when multiple candidates were available. This preprocessing step reduced vocabulary sparsity and improved token consistency prior to subword segmentation.

Subsequently, named entities were extracted using the in-house developed Bi-LSTM–based POS tagger (Nathani et al., 2023). The extracted named entities were then classified into the MUC-6 categories (Grishman et al., 1996) through a rule-based approach. These annotated entities were cross-referenced with a knowledge base containing target language translations for source language organization and location names.

The Named Entity Translation module relied on resources compiled from multiple sources. Gazetteers of Indian locations and organization names were collected from publicly available repositories such as the Wikipedia category lists, and official government publications. Entities not present in the knowledge base were transliterated using a rule-based transliteration scheme, which maps source graphemes to phoneme-equivalent representations before rendering them in the target script. This approach preserved phonological similarity across languages, ensuring that named entities remained intelligible even in the absence of dictionary support. In future work, we plan to explore phoneme-to-grapheme transliteration models trained using transformer-based sequence-to-sequence architectures for improved accuracy.

A rule-based NER system was employed to extract named entities from the source language corpus (Suri et al., 2024). Once identified, these entities were searched in the knowledge base for their corresponding English translations. When a translation was available, the entity in the source language corpus was replaced with its target language equivalent. In cases where no translation was present in the knowledge base, the entities were instead transliterated into the target language and then replaced in the corpus.

This process constituted the Named Entity Translation module, which systematically identified named entities and translated or transliterated them into target language, depending on the availability of translations (Sharma et al., 2023; Joshi & Katyayan, 2023). The functioning of this module is illustrated in Figure 1.

A BiLSTM-based POS tagger was used to bootstrap named entity recognition, as gold-standard BIO-annotated NER corpora for Manipuri and Bodo were not available at the time of system development. POS tags provided coarse-grained syntactic cues (e.g., proper noun categories) which, when combined with rule-based heuristics, enabled the identification of named entities. While this approach proved effective in resource-constrained settings, we acknowledge its limitations compared to transformer-based BIO tagging. Future extensions of this work will evaluate pre-trained multilingual NER models such as IndicNER and

XLM-R–based fine-tuned taggers, which are expected to improve robustness.

3.2 Sub-Wording

We adopted Byte Pair Encoding (BPE) for subword segmentation using SentencePiece. BPE was selected due to its effectiveness in balancing vocabulary compactness and coverage, a crucial factor for low-resource settings where unseen tokens are frequent (Sennrich et al., 2016). Although alternative segmentation schemes such as the Unigram Language Model (Kudo & Richardson, 2018) have been shown to perform competitively in recent work (Ahmed et al., 2023; WMT 2023 Shared Task Report), preliminary trials indicated that BPE produced more stable vocabularies across our diverse Indic language pairs. A systematic comparison with alternative subword models remains an avenue for future research.

3.3 Training the Model

For the training of the NMT systems, preprocessing steps were applied. This process was as follows: Part-of-Speech (POS) tagging was first applied to source language sentences, after which Named Entity Recognition (NER) was conducted using a rule-based module. The identified named entities were then translated or transliterated according to the procedure described in the previous section, thereby producing an augmented source sentence for the training corpus.

For example, consider the Sindhi sentence: "নিশীথ জোশী নতুন দিল্লী ৰ ইন্দিৰা গান্ধী আন্তর্জাতিক বিমানবন্দৰ পৰা জয়পুৰলৈ যাত্রা কৰি আছিল।" In this sentence, "নিশীথ জোশী (Person), নতুন দিল্লী (Location), জয়পুৰ (Location)," and "ইন্দিৰা গান্ধী আন্তর্জাতিক বিমানবন্দৰ (Organization)" are named entities. Among these, "নিশীথ জোশী" and "জয়পুৰ" were not present in the knowledge base and were therefore transliterated as "Nisheeth Joshi" and "Jaipur." The remaining entities were looked up in the knowledge base sequentially. While "নতুন দিল্লী" was not found and was thus transliterated as "New Delhi," "ইন্দিৰা গান্ধী আন্তর্জাতিক বিমানবন্দৰ" was translated as "Indira Gandhi International Airport."

Using this methodology, the entire training corpus was augmented with translated and transliterated named entities. The workflow of the system is illustrated in Figure 2, while the hyperparameters used for training both systems are summarized in Table 1. The overall approach is named as HEMANT (Highly Efficient Machine Assisted Natural Translation).

Our systems were built on top of the No Language Left Behind (NLLB-200) pretrained multilingual model (Costa-jussà et al., 2022), which supports several Indic languages including Assamese. Instead of training from scratch—which is often infeasible for low-resource settings due to limited parallel corpora—we adopted a transfer learning approach by fine-tuning NLLB-200 on the shared task training data provided by the organizers.

Parameter	Value
No. of Encoding Layers	6
No. of Decoding Layers	6
Early Stopping	
metric	bleu
min_improvement	0.2
steps	6
Optimizer	Adam
beta_1	0.8
beta_2	0.998
learning_rate	1.0
droupout	0.25
Regularization	
type	l1_l2
scale	1e-4
Minimum_learning_rate	0.00001
Max_steps	1000000

Tabel 1: Hyperparameters Used in Training NMT Models

The base architecture of NLLB-200 is a Transformer encoder–decoder model with multi-head attention, residual connections, and layer normalization, optimized for cross-lingual transfer. Fine-tuning was carried out by unfreezing all layers of the model, while maintaining pretrained multilingual subword embeddings. Subword embeddings were initialized from the NLLB model's vocabulary, which itself is trained using a combination of BPE and SentencePiece across 200 languages. This initialization ensured robust handling of rare tokens and morphologically complex forms.

Fine-tuning was performed separately for each language pair, with learning rate schedules tuned to prevent catastrophic forgetting of pretrained knowledge. For language pairs with extreme data scarcity (e.g., Bodo), multi-directional fine-tuning

was explored by jointly optimizing the model on related language pairs, leveraging cross-lingual similarity between Assamese and Bengali for Indo-Aryan, and Manipuri and Bodo for Tibeto-Burman families.

This strategy balanced the benefits of pretrained multilingual representations with task-specific adaptation. We found that fine-tuning NLLB-200 significantly stabilized training compared to Transformer models trained from scratch, which often struggled to converge on the limited IndicMT corpora.

4 Evaluation

We participated in the shared task by training the models on the training corpus provided by the organizers and submitted the outputs generated by the systems, using the test corpus, for official evaluation. The corresponding results provided by the organizing team are presented in Table 2. For Assamese-English, English-Assamese, English-Manipuri and Manipuri-English language pairs; our system performed fairly well. We could not provide the same results for English-Bodo, possibly due to very less training corpus.

5 Conclusion

This paper presented HEMANT, our submission to the WMT 2025 Indic MT shared task, focusing on five low-resource language pairs. The integration of a Named Entity Translation module reduced out-of-vocabulary errors and improved translation fluency. While absolute scores remain modest, the relative improvements highlight the value of linguistically informed preprocessing.

In future work, we plan to explore cross-lingual transfer strategies by leveraging related languages (e.g., Bengali–Assamese). Compare alternative subwording methods (BPE vs Unigram LM). Incorporate backtranslation and synthetic data augmentation for extremely low-resource languages and replace POS-based NER heuristics with multilingual pretrained BIO tagging models.

Acknowledgments

This work is supported by the funding received from the Ministry of Electronics and Information Technology, Government of India for the project "English to Indian Languages and vice versa Machine Translation System" under National Language Translation Mission (NLTM): Bhashini through administrative approval no. 11(1)/2022-HCC(TDIL) Part 5 and funding received from Department of Science and Technology, Government of India through grant number DST/SHRIC/SHRI-24/2023 for project entitled, "Bidirectional Dhundhari-Hindi Machine Translation System".

References

Ahmed, T., Hasan, M. K., Hoque, M. T., & Sultana, N. (2023). A comparative study on subword segmentation strategies for low-resource neural machine translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 912–920). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.87

Ahuja, K., Dandapat, S., Dave, S., Khapra, M. M., Kumar, P., & Shrivastava, M. (2022). IndicTrans: An open-source model for transliteration and translation for Indic languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 127–137). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-demo.14

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & Guzmán, F. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*. https://doi.org/10.48550/arXiv.2207.04672

Grishman, R., & Sundheim, B. M. (1996). Design of the MUC-6 evaluation. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996* (pp. 413–422). Association for Computational Linguistics.

Joshi, N., & Katyayan, P. (2023). Improving English-Bharti Braille machine translation through proper name entity translation. In *Proceedings of the 3rd International Conference on ICT for Digital, Smart, and Sustainable Development (ICIDSSD 2022)* (pp. 168–174). European Alliance for Innovation.

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 66–71). Association for Computational Linguistics.

Nathani, B., Arora, P., Joshi, N., Katyayan, P., Rathore, S. S., & Dadlani, C. P. (2023). Sindhi POS tagger using LSTM and pre-trained word embeddings. In *XVIII International Conference on Data Science

and Intelligent Analysis of Information* (pp. 37–45). Springer Nature Switzerland.

Nguyen, T. Q., & Chiang, D. (2017). Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)* (pp. 296–301). Association for Computational Linguistics. https://aclanthology.org/I17-1050

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics.

Sharma, R., Katyayan, P., & Joshi, N. (2023). Improving the quality of neural machine translation through proper translation of name entities. In *2023 6th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1–4). IEEE.

Suri, D., Malviya, N., & Joshi, N. (2024). Rule-based named entity recognition for Hindi. In *International Conference on Artificial Intelligence and Speech Technology* (pp. 55–62). Springer Nature Switzerland.

WMT 2023 Shared Task Report. (2023). Findings of the WMT 2023 shared tasks on machine translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT 2023)* (pp. 1–45). Association for Computational Linguistics. https://aclanthology.org/2023.wmt-1.0

Zoph, B., & Knight, K. (2016). Multi-source neural translation. In *Proceedings of NAACL-HLT 2016* (pp. 30–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1004

1

Language Pair	BLEU	Meteor	ROUGE-L	chrf	TER
As-En	14.90698584	0.6152953104	0.6125623657	60.29148355	71.32659784
En-As	1.810109903	0.05849541589	0.003	27.45025141	98.66174223
En-Mni	4.15145205	0.1464458762	0.009866666667	41.43192248	89.59891851
Mni-En	3.056814809	0.2212412906	0.2506661424	35.61296964	139.025647
En-Bod	1.350078456	0.04016354522	0.1678	17.05109642	106.111629

Tabel 2: Evaluation Results

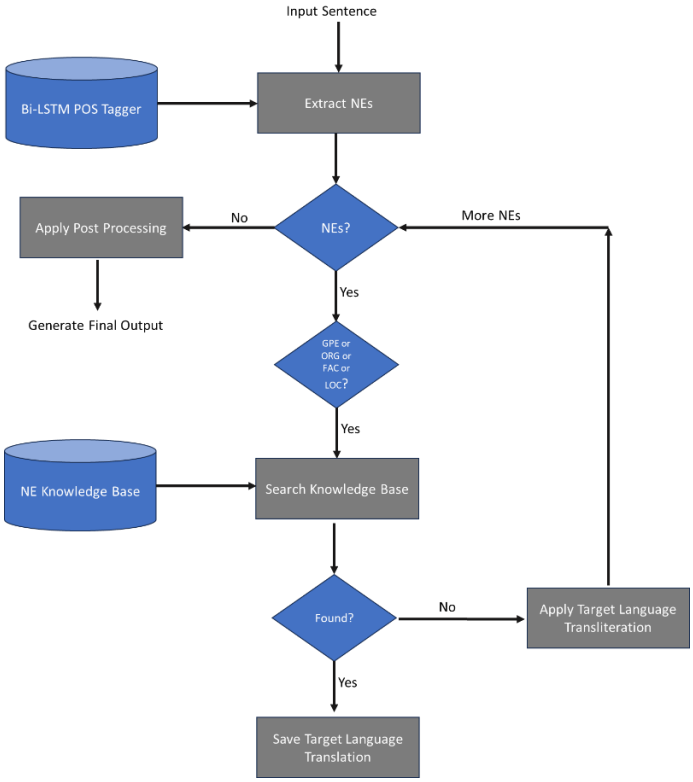


Figure 1: Named Entity Translation Module

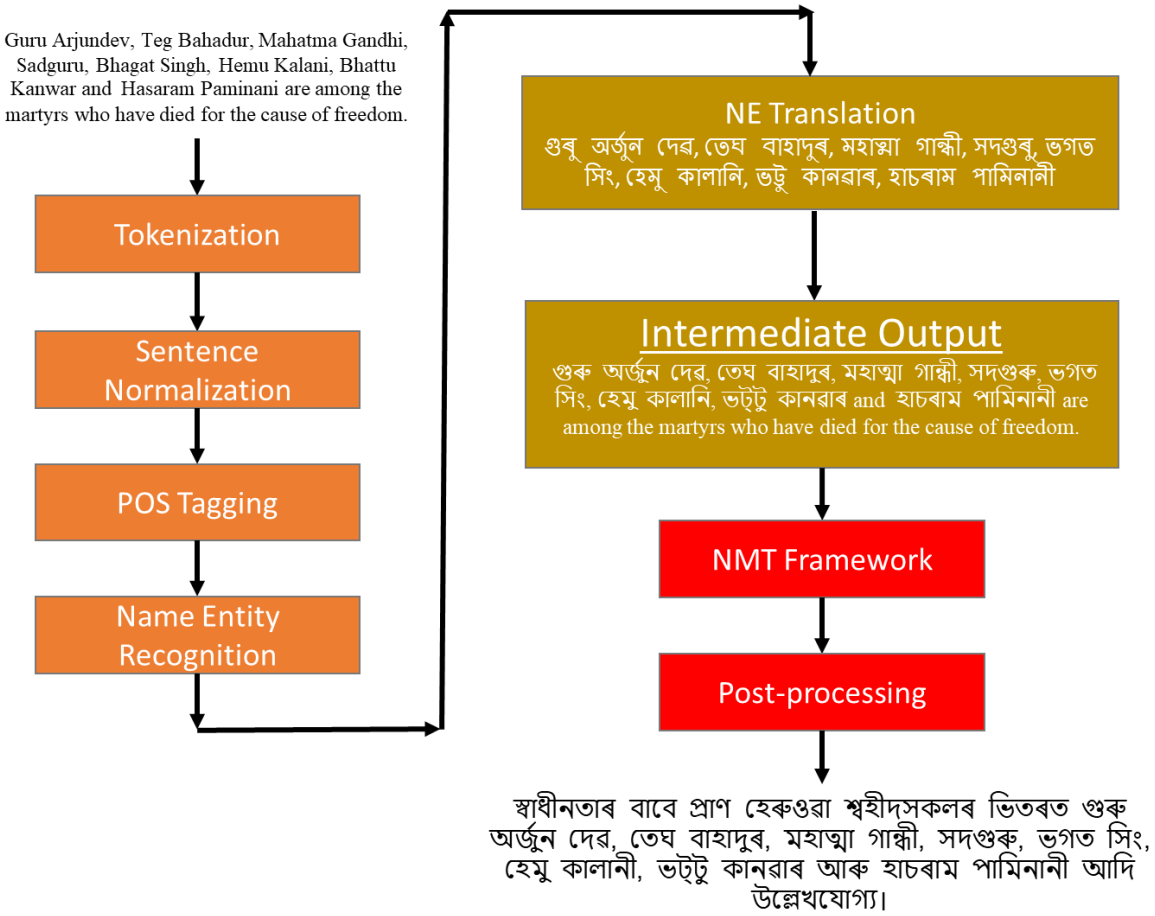


Figure 2: HEMANT Approach

TranssionMT's Submission to the Indic MT Shared Task in WMT 2025

ZeBiao Zhou and Hui Li and XiangXun Zhu and KangZhen Liu
Transsion, Shenzhen, Guangdong, China

Abstract

This study tackles the WMT 2025 low-resource Indic language translation task (EnglishAssamese, EnglishManipuri) by proposing a cross-iterative back-translation and data augmentation strategy using dual pre-trained models.Leveraging IndicTrans2_1B and NLLB_3.3B, the approach alternates fine-tuning and back-translation to iteratively generate high-quality pseudo-parallel corpora. Monolingual data relevance is enhanced via semantic similarity filtering with all-mpnet-base-v2, while training data is cleaned and normalized to improve quality. During inference, outputs from both fine-tuned models are combined to further boost translation performance in low-resource scenarios.

1 Introduction

India boasts a rich linguistic ecosystem, yet numerous languages suffer from limited digital resources. These low-resource languages face significant challenges in the construction and application of machine translation systems. Particularly, Assamese and Manipuri in the northeastern region not only lack parallel corpora and high-quality monolingual corpora but also exhibit large differences in linguistic structure and writing systems, posing additional difficulties for the training of Neural Machine Translation (NMT) models. Under low-resource conditions, traditional neural machine translation methods cannot fully leverage the advantages of large-scale data, resulting in limited model generalization ability and translation quality. Therefore, exploring how to efficiently utilize limited bilingual and monolingual data and effectively transfer cross-lingual knowledge has become a core issue in improving the translation performance of low-resource languages. The study selects open-source IndicTrans2-1B and NLLB-3.3B as the core translation models, combines multiple rounds of iterative back-translation to generate high-quality pseudo-parallel corpora, uses semantic similarity filtering technology to enhance the alignment between monolingual data and the task, and reduces the interference of noisy data on training through strict data cleaning and standardization operations. During the inference phase, the outputs of the two models are compared and selected, with the optimal result serving as the final translation output. This study aims to verify the effectiveness of cross-model collaborative back-translation mechanisms, data similarity augmentation, and multi-source result fusion in low-resource translation tasks, providing reusable technical routes and empirical experience for future multilingual low-resource machine translation research.

2 Dataset

All parallel data used in this study are derived from the official bilingual data provided by the WMT 2025 Low-Resource Indic Language Translation track, covering two directions: EnglishAssamese (en-as) and EnglishManipuri (en-mni). The data scale is shown in Table 1. Among them, the en-as direction contains 54,000 training sentence pairs, the en-mni direction contains 23,000 training sentence pairs, and both directions provide validation sets and test sets respectively.

A sampling analysis of the official test set reveals a concentrated domain distribution: healthcare accounts for 65.29%, entertainment and sports for 23.56%, and culture for 11.15%. To maximize domain consistency with the test set during the data augmentation phase, the study collects English monolingual data from the NLLB open corpus, BPCC open-source dataset, and specific website crawls. The open-source semantic similarity model all-mpnet-base-v2 is used to calculate the semantic similarity between the collected data and the test set samples. Sampling and filtering are performed in high-similarity data according to the

Proceedings of the Tenth Conference on Machine Translation, pages 1271–1275

Language Pair	Train	Val	Test
en-as	54,000	2,000	2,000
en-mni	23,000	1,000	1,000

Table 1: Scale of WMT 2025 Official Bilingual Dataset

above domain proportions, ultimately obtaining approximately 100,000 highly relevant English monolingual sentences for back-translation to generate pseudo-parallel corpora.

During the data cleaning phase, strict processing is uniformly applied to bilingual and monolingual data: removing sentences containing URLs, HTML tags, and non-linguistic characters; eliminating samples that failed to be translated or deviated from the source language in back-translation; standardizing symbols, checking and correcting English capitalization rules for the first letter; and removing duplicate sentences and abnormally short sentences. These operations significantly reduce the proportion of noisy data and ensure that the data domain distribution is highly consistent with the official test set, providing high-quality data support for subsequent cross-iterative back-translation and model optimization.

3 System Methodology

3.1 Pre-trained Models

This study is based on two open-source multilingual neural machine translation pre-trained models: IndicTrans2_1B (Kunchukuttan et al., 2023) and NLLB_3.3B (Fan et al., 2022).

• **IndicTrans2_1B**: A Transformer-based machine translation model optimized for 22 official languages of India and various related languages. It performs excellently in many-to-many, many-to-one, and one-to-many translation tasks, especially suitable for handling Indic languages with complex morphology and scarce training data (Kunchukuttan et al., 2023).

• **NLLB_3.3B (No Language Left Behind)**: A large-scale multilingual translation model proposed by Meta AI, covering more than 200 languages and possessing strong generalization ability in cross-lingual transfer (Fan et al., 2022).

The core reason for selecting these two models lies in their complementarity in multilingual environments: IndicTrans2_1B has obvious advantages in the fine-grained processing of Indic languages, while NLLB_3.3B is more robust in cross-lingual structure mapping and low-resource direction generalization. Their combination helps obtain more diverse and high-quality pseudo-parallel data under extremely low-resource conditions.

3.2 Direction-Specific Fine-Tuning

In the first phase of system construction, the above two models are respectively fine-tuned in one-to-one directions on the bilingual parallel corpora provided by WMT 2025, covering four translation directions: en→as, as→en, en→mni, and mni→en.

One-way translation fine-tuning at the granularity of translation directions enables the model to focus on learning the syntactic, lexical, and domain features of that direction. Compared with directly training a multilingual multi-directional model, it can avoid cross-direction interference and achieve higher convergence speed and better direction adaptability in low-resource scenarios. The training results of this phase serve as the baseline models for subsequent back-translation augmentation.

During the fine-tuning phase for NLLB_3.3B, LoRA (Low-Rank Adaptation) parameter-efficient fine-tuning technology is adopted, with specific configurations as follows (Hu et al., 2021):

• rank: 128
• alpha: 256
• dropout: 0.1
• Fine-tuning modules: All linear layers

LoRA injects low-rank matrix parameters into the model's linear layers, keeping most original parameters frozen and only updating a small number of trainable parameters, which significantly reduces memory usage and training costs while maintaining model performance. This design is particularly effective for models at the 3.3B scale, enabling high-quality directional fine-tuning to be completed under single-card or low-resource computing power conditions (Hu et al., 2021).

3.3 Monolingual Data Back-Translation Augmentation

The second phase introduces 100,000 English monolingual sentences with a domain proportion highly consistent with the test set to improve the model's adaptability in the target domain. This monolingual data is sourced from the NLLB open corpus, BPCC open-source data, and domain-specific web crawls. It is matched and filtered with test set samples using the all-mpnet-base-v2 semantic similarity model to ensure the domain distribution proportion is consistent with the test

set (65.29% healthcare, 23.56% entertainment and sports, 11.15% culture).

Based on the two fine-tuned models obtained in the first phase, dual-model back-translation is implemented, which is a widely used data augmentation technique in low-resource machine translation to generate pseudo-parallel corpora (Sennrich et al., 2016):

1. IndicTrans2_1B translates English monolingual sentences into the target language, generating pseudo-parallel corpus set D1;

2. NLLB_3.3B translates English monolingual sentences into the target language, generating pseudo-parallel corpus set D2.

D1 and D2 are respectively merged with the official parallel data to fine-tune IndicTrans2_1B and NLLB_3.3B again, forming the first-round augmented models. Taking "back-translation → merging → fine-tuning" as a cycle, the iteration is performed until the BLEU score of the development set no longer improves. In actual experiments, significant improvements can be achieved with two iterations, and the third iteration is difficult to bring additional benefits. Therefore, two iterations are finally adopted as the optimal solution.

This dual-model iterative back-translation augmentation method fully leverages the complementary advantages of the two pre-trained models in language modeling and cross-lingual generalization, significantly enriches the diversity and domain coverage of training data in low-resource directions, and thereby improves the translation performance of the final system (Sennrich et al., 2016).

4 Experimental Results and Analysis

Table 2 shows the BLEU score performance of different systems and data augmentation strategies in the four translation directions (en→as, en→mni, as→en, mni→en). The experiment compares the performance changes of IndicTrans2_1B and NLLB_3.3B under one-way fine-tuning on official data, different back-translation data augmentations, and dual-model iterative back-translation.

Strategy	**en→as**	**en→mni**
IndicTrans2-1B	16.33	10.28
+OFT-off	23.80	16.24
+OFT-off+BT-it	–	–
+OFT-off+BT-nllb	–	–
+OFT-off+BT-itnllb	25.92	24.34
NLLB-3.3B	17.04	15.01
+OFT-off	24.52	21.29
+OFT-off+BT-it	29.72	25.19
+OFT-off+BT-nllb	28.32	25.28
+OFT-off+BT-itnllb	30.61	27.71
+OFT-off+DBT (P2)	32.11	28.92
Strategy	**as→en**	**mni→en**
IndicTrans2-1B	29.20	34.74
+OFT-off	40.36	44.35
+OFT-off+BT-it	40.10	43.86
+OFT-off+BT-nllb	41.61	44.56
+OFT-off+BT-itnllb	–	–
NLLB-3.3B	30.88	30.77
+OFT-off	37.69	37.75
+OFT-off+BT-it	–	–
+OFT-off+BT-nllb	–	–
+OFT-off+BT-itnllb	–	–
+OFT-off+DBT (P2)	–	–

Table 2: BLEU Scores of Different Systems and Data Augmentation Strategies on WMT 2025 Development Set. Note: "–" indicates that this combination was not tested in this direction or the results were not included in the statistics. We use the following abbreviations: OFT denotes One-way Fine-Tuning, off denotes official data,BT denotes Back-Translation, it and nllb denote different back-translated datasets, LoRA denotes Low-Rank Adaptation, and DBT denotes Dual Back-Translation.

4.1 Significant Gains from One-Way Fine-Tuning

After one-way (one-to-one) fine-tuning using official parallel corpora, both baseline models show significant improvements in BLEU scores across all tested translation directions:

• IndicTrans2_1B increases from 29.20 to 40.36 (+11.16 BLEU) in the as→en direction, and from 34.74 to 44.35 (+9.61 BLEU) in the mni→en direction;

• NLLB_3.3B (LoRA fine-tuning) rises from 17.04 to 24.52 (+7.49 BLEU) in the en→as direction, and from 15.01 to 21.29 (+6.28 BLEU) in the en→mni direction.

The above results demonstrate that one-way fine-tuning can effectively reduce multi-task interference and improve translation quality in specific directions under low-resource conditions.

4.2 Single-Model Back-Translation Augmentation Effect

When introducing the first round of back-translation augmentation, model performance continues to improve, but the effect depends on the source of back-translated data:

• IndicTrans2_1B: Using back-translated data from NLLB_3.3B (41.61 BLEU in the as→en direction) is superior to using its own back-translated data (40.10 BLEU); a slight gain is also maintained in the mni→en direction (44.56 vs. 43.86);

• NLLB_3.3B (LoRA): Using back-translated data from IndicTrans2_1B (29.72 BLEU in the en→as direction) is better than self-back-translated data (28.32 BLEU); the performance in the en→mni direction is close (25.19 vs. 25.28).

This indicates that pseudo-parallel data generated across models has complementarity in syntactic and lexical distributions, which can reduce noise accumulation in self-back-translation.

4.3 Dual-Model Back-Translation and Iterative Optimization

After adding the back-translated data of both models to the training simultaneously (dual-model back-translation), NLLB_3.3B (LoRA) achieves a BLEU score of 30.61 in the en→as direction and 27.71 in the en→mni direction; further performing the second round of dual back-translation iteration results in 32.11 BLEU in the en→as direction (+1.50 compared to the previous stage) and 28.92 BLEU in the en→mni direction (+1.21). The results show that multiple rounds of back-translation can bring additional benefits, but the marginal gain diminishes.

4.4 Value of LoRA Fine-Tuning

Considering the 3.3B parameter size of NLLB_3.3B, this study adopts LoRA (rank=128, alpha=256, dropout=0.1, injected into fully connected layers) for efficient one-way fine-tuning. Under the premise of low memory usage, a significant BLEU improvement is still achieved, making multi-stage data augmentation possible under limited computing power conditions (Hu et al., 2021).

5 Conclusion

This study addresses the low-resource translation task by combining two complementary multilingual pre-trained models, IndicTrans2_1B and NLLB_3.3B, and proposes a system construction method of one-way fine-tuning for specific translation directions and dual-model iterative back-translation augmentation. The introduction of LoRA parameter-efficient fine-tuning technology on NLLB_3.3B significantly reduces memory and computational costs, enabling multi-stage data augmentation under limited computing power conditions.

Experimental results show that:

1. One-way fine-tuning can significantly improve BLEU scores in low-resource translation directions (up to +11.16 BLEU), effectively reducing multilingual multi-directional interference;

2. Cross-model back-translation data augmentation is superior to single-model self-back-translation, proving that pseudo-parallel data generated by different models has complementarity in syntactic and lexical distributions;

3. Dual-model back-translation + multi-round iteration can further improve model performance, although the gain tends to converge after the second round;

4. LoRA technology balances efficiency and effectiveness in the directional fine-tuning of ultra-large-scale models, enabling the performance of low-resource translation directions to approach the improvement range of full fine-tuning (Hu et al., 2021).

Overall, the system method in this study fully leverages the complementary advantages of the two pre-trained models, combines parameter-efficient fine-tuning and dual-model iterative back-translation, and achieves significant BLEU improvements in the WMT 2025 low-resource task, providing a feasible and efficient reference scheme for the construction of low-resource machine translation systems. Future work will further explore adaptive back-translation data screening for multi-

model collaboration and the introduction of multimodal auxiliary information in low-resource scenarios to break through performance bottlenecks.

6 Future Work

On the basis of improving the low-resource translation performance achieved in this study, future work will continue to expand in the following two directions:

6.1 In-depth Utilization of Monolingual Data

Although parallel corpora for low-resource languages are limited, monolingual texts are often relatively abundant. Future work will consider:

1. Continual Monolingual Pretraining: Conducting continuous training on existing models (such as IndicTrans2_1B, NLLB_3.3B) using a large amount of Indic monolingual data to improve language fluency and localized expression ability;

2. Denoising Self-Supervised Training: Drawing on methods such as mBART and MASS, enabling the model to better grasp contextual dependencies and syntactic structures through tasks such as Masked Span Prediction and Noising & Reconstruction;

3. Combining Monolingual Back-Translation and Forward Translation: Constructing bidirectional pseudo-parallel data by combining monolingual data, that is, adding forward translation data generated from the target language to the source language on the basis of back-translation, to further improve the model's generalization ability.

6.2 Application of Large Language Models in Translation

With the development of multilingual Large Language Model (LLM) capabilities, introducing them into low-resource translation tasks has potential. Future work will consider:

1. LLM-as-Translator: Using general-purpose LLMs (such as Qwen, LLaMA, Mixtral, mT5) for direct translation or back-translation to generate higher-quality pseudo-parallel data that is more contextually appropriate;

2. Parameter-Efficient Fine-Tuning (PEFT) for Small Languages: Quickly adapting LLMs to specific small languages and domains through methods such as LoRA, Prefix Tuning, and Adapters, reducing computational costs while improving performance in low-resource scenarios;

3. Multi-Task Learning and Instruction-Tuning: Simultaneously training tasks such as translation, question answering, and paraphrasing on LLMs, and improving their ability to understand and generate low-resource languages through multi-task transfer effects.

References

Angela Fan and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Anoop Kunchukuttan and 1 others. 2023. Indictrans2: Towards high-quality and low-resource machine translation for indic languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Laniqo at WMT25 Terminology Translation Task: A Multi-Objective Reranking Strategy for Terminology-Aware Translation via Pareto-Optimal Decoding

Kamil Guttmann[1,2], **Adrian Charkiewicz**[1,2], **Zofia Rostek**[1], **Mikołaj Pokrywka**[1,2], **Artur Nowakowski** [1,2]
[1] Laniqo, Poznań, Poland
[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland
{name}.{surname}@laniqo.com

Abstract

This paper describes the Laniqo system submitted to the WMT25 Terminology Translation Task. Our approach uses a Large Language Model fine-tuned on parallel data augmented with source-side terminology constraints. To select the final translation from a set of generated candidates, we introduce Pareto-Optimal Decoding – a multi-objective reranking strategy. This method balances translation quality with term accuracy by leveraging several quality estimation metrics alongside Term Success Rate (TSR). Our system achieves TSR greater than 0.99 across all language pairs on the Shared Task testset, demonstrating the effectiveness of the proposed approach.

1 Introduction

The shared task consists of two tracks: sentence-level translation and document-level translation. Our submission focuses exclusively on the sentence-level track, where each sentence is translated independently using provided terminology constraints.

The shared task requires systems to be evaluated in three distinct modes:

- **No Terminology (`noterm`)**: The system is only provided with the input sentences.
- **Proper Terminology (`proper`)**: The system receives the input text along with a dictionary of domain-specific terminology pairs.
- **Random Terminology (`random`)**: The system is provided with the input text and a dictionary of randomly sampled terms from the source and target texts.

Our system builds upon the `EuroLLM-9B-Instruct` model[1] (Martins et al., 2025), which served as the baseline for our experiments in Terminology-Aware Machine Translation (MT). While this model already offers multilingual translation capabilities, it is not explicitly designed to handle translation with domain-specific terminology constraints. To further improve terminology control and translation quality, we explored several complementary strategies:

1. **Source-side terminology replacement**: Before translation, we replaced source-language terms with their corresponding target-language equivalents. This was combined with explicit prompts designed to guide the model in retaining or correctly adapting the inserted terms.
2. **Fine-tuning on glossary-augmented data**: We fine-tuned the base model on parallel data augmented with terminology automatically aligned between source and target segments. The objective of this training was to expose the model to code-switched source sentences, enabling it to learn the mechanism for incorporating provided target-language terms during translation.
3. **Pareto-Optimal Decoding**: We propose a reranking strategy that integrates multiple reference-free Quality Estimation (QE) metrics along with terminology accuracy to select the most accurate and terminology-compliant translation candidate.

In addition, we investigated prompt engineering techniques, such as structured instructions and two-shot examples, aimed at improving system robustness.

The methods described above proved effective in both improving the correct use of terminology and maintaining overall translation quality, as confirmed by automatic metrics.

[1] https://huggingface.co/utter-project/EuroLLM-9B-Instruct

Proceedings of the Tenth Conference on Machine Translation, pages 1276–1283
November 8-9, 2025

2 Related Work

Handling specialized terminology in Neural Machine Translation (NMT) has received considerable research interest in recent years. The proposed methods can be broadly classified into three main categories:

- **Constrained decoding**: These methods modify the beam search algorithm by restricting the search space to ensure that the generated hypotheses include the specified terminology (Hokamp and Liu, 2017; Nowakowski and Jassem, 2021). Furthermore, the application of negative constraints has been studied for re-translating sentences where an initial translation failed to incorporate the required terms (Bogoychev and Chen, 2023).

- **Placeholding**: This approach involves replacing source terms with special placeholder tokens (Michon et al., 2020). The model is then trained to copy these placeholders into the hypothesis, enabling the target terms to be injected during post-processing. The main disadvantage of this method is that masking the source terms can result in a lack of context, which can lead to a degradation in fluency of the final translation.

- **Source Text Constraints**: This technique involves augmenting the training data with inline terminology constraints (Dinu et al., 2019; Bergmanis and Pinnis, 2021). The source text is augmented by adding the target term alongside its equivalent in the source language. These terms are typically annotated with source factors (Sennrich and Haddow, 2016) or pre-defined tags. This approach has proven particularly effective for morphologically rich languages when combined with Target Lemma Annotations (Bergmanis and Pinnis, 2021). The model generates the appropriate morphological form of the target term, ensuring that it adheres to the grammatical rules of the target language.

The emergence of multilingual Large Language Models (LLMs), such as EuroLLM (Martins et al., 2025) and Tower+ (Rei et al., 2025), represents a significant shift in MT towards the LLM era (Kocmi et al., 2024). The ability of LLMs to follow natural language instructions embedded within a prompt opens up new possibilities for adhering to terminology constraints.

Several possible strategies that leverage this capability have been explored. One approach involves using LLMs to generate term-rich synthetic data for fine-tuning traditional NMT models (Moslem et al., 2023b). Another line of work uses LLMs for automatic post-editing, prompting the model to inject missing terms into an existing translation (Bogoychev and Chen, 2023). A third, more direct method involves providing the LLM with the source text and a list of terminology constraints which must be included in the output (Moslem et al., 2023a).

Our approach builds on this direct prompting method by extending it through the integration of constraints in the source text.

While the primary objective of the Shared Task is to ensure the correct translation of specified terminology, maintaining the overall quality of the translation remains a critical factor. Well-established methods for improving the quality of machine translation, such as Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004) and QE reranking, have consistently demonstrated their effectiveness in various research studies in recent years (Nowakowski et al., 2022; Finkelstein and Freitag, 2024; Guttmann et al., 2024). However, these approaches are typically designed to optimize for a single metric.

This task requires the simultaneous optimization of two distinct aspects: term accuracy and general translation quality. To address this multi-objective problem, we propose a method that balances these potentially conflicting criteria.

3 Approach

3.1 Replace Method

Given a source segment and a glossary containing terminology pairs, we replaced each source term in the input text with its corresponding target-language term (see Table 1 for example). In contrast to previous works (Nieminen, 2023; Ri et al., 2021), where target terminology was appended to or replaced within the source text and subsequently enclosed by special tags, our approach, similar to the data augmentation method proposed by Song et al. (2019), directly replaces terms with their target-language equivalents without introducing any additional markup. Directly inserting target

```
<|im_start|>system
You are a professional {src_lang} to {tgt_lang} translator.
Your goal is to accurately convey the meaning and nuances of the
original {src_lang} text while adhering to {tgt_lang} grammar,
vocabulary, and cultural sensitivities.
<|im_end|>
<|im_start|>user
Some words have been pre-translated. You may need to correct them
in the final translation for a better fit into the context.
{src_term1} -> {tgt_term1}
{src_term2} -> {tgt_term2}
{src_term3} -> {tgt_term3}

Translate the following {src_lang} source text to {tgt_lang}:
{src_lang}: {src_text_replaced}
{tgt_lang}: <|im_end|>
<|im_start|>assistant
```

Listing 1: Baseline terminology-aware translation prompt.

language terms into the source sentence results in code-switching, enabling the model to adapt the grammatical form of the target terms to fit them into the sentence structure during inference.

Source	"In the Switch **Data Provider** dialog:"
Terminology	{"data provider": "Daten-provider"}
Replaced	"In the Switch **Datenprovider** dialog:"

Table 1: Example of terminology handling method applied to the source sentence.

The replacement process involved lemmatizing both the source text and the glossary terms using the `simplemma` library (Barbaresi, 2021). Each source term was matched against the lemmatized input, and when a match was found, it was substituted with the corresponding target term.

3.2 Prompt

The base prompt was designed to ensure the integration of glossary terms. Since replacing source terms with their target-language equivalents can obscure grammatical cues needed for correct inflection, the prompt also includes a list of original source phrases alongside their corresponding target terms. This provides the model with additional context, helping it resolve potential ambiguities and adjust terminology to the surrounding syntax when necessary. The full prompt is shown in Listing 1.

During the experiments, we observed that a large proportion of the mistranslated terminology consisted of phrases that were identical or very similar in both the source and target languages - differing only in casing. When the model received a text in which a term had been replaced with an identical or nearly identical term in the target language, it did not recognize the change and consequently translated it into an equivalent term in the target language. We conducted additional experiments, making some adjustments to the prompt to draw the model's attention to this issue. As a result, we added the following instruction to the prompt: `(keep already translated {tgt_lang} words - {tgt_terms})`, for instance `Translate the following English source text to Spanish (keep already translated Spanish words - desea, Decida)`. This solution improved translation quality, leading to more frequent and correct usage of the specified terminology.

3.3 Few-Shot Prompting

We also conducted experiments on few-shot prompting. We found that providing two translation examples with terminology (in the appropriate source and target languages) improves translation quality, both in general translation metrics and in term accuracy, averaged across all three language pairs.

Parameter	Value/Description
LoRA Configuration	
LoRA Rank (r)	16
LoRA Alpha (α)	32
LoRA Dropout	0.15
General Training Configuration	
Per-Device Batch Size	4
Gradient Accumulation Steps	8
Learning Rate	1×10^{-4}
LR Scheduler Type	Inverse Square Root

Table 2: Fine-tuning hyperparameters

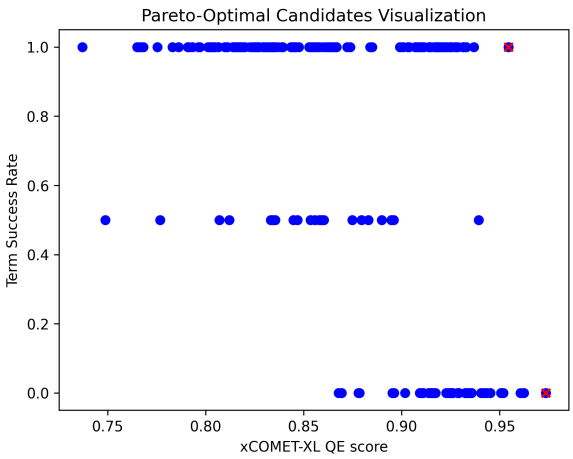


Figure 1: Visualization of Pareto-Optimal Decoding for a sample sentence. Each blue point represents one of 100 translation candidates. The red $\times$ markers highlight the non-dominated solutions. The top marked solution would be chosen by our algorithm to maximize the TSR score.

3.4 Terminology Aware Fine-Tuning

We fine-tuned the `EuroLLM-9B-Instruct` model for English $\rightarrow$ German, English $\rightarrow$ Spanish, and English $\rightarrow$ Russian terminology translation tasks. For each language pair, we used a training set of 200,000 sentence pairs randomly sampled from the OPUS corpora (Tiedemann and Nygaard, 2004).

To prepare data, we adopted a similar approach to the Target Lemma Annotations (TLA) method proposed in (Bergmanis and Pinnis, 2021). Specifically, we selectively sampled nouns and verbs from the target sentences via POS-tagging using `Stanza` (Qi et al., 2020). These words were then aligned with their corresponding source terms using the `fast_align` tool (Dyer et al., 2013), creating parallel term pairs for each sentence. These sentences and term pairs were then formatted using the prompt template shown in Listing 1.

Fine-tuning was performed using LoRA (Hu et al., 2021) over 3 epochs on 4$\times$A100 GPUs conducted through the Oumi (Oumi Community) framework. The specific training hyperparameters are detailed in Table 2.

3.5 Pareto-Optimal Decoding

In the final stage, we used epsilon sampling with $\epsilon = 0.02$ and $T = 1$ to generate 100 candidate translations for each source sentence. This method has previously been found to be effective for creating diverse samples for techniques such as MBR decoding and QE reranking (Freitag et al., 2023). Next, we scored each source-candidate pair using several QE metrics, namely `xCOMET`[2] (Guerreiro et al., 2024), `ReMedy`[3] (Tan and Monz, 2025) and `MetricX`[4] (Juraska et al., 2024). Additionally, we calculated the Term Success Rate (TSR) (Semenov et al., 2023) by verifying the presence of the lemmatized source term words within the lemmatized candidate translation sentence.

Previous research has shown that, while methods such as QE reranking and MBR decoding significantly improve translation quality, they can lead to overfitting to the utility metric (Pombal et al., 2025). We anticipate that simply reranking according to TSR could result in a substantial decrease in translation quality, particularly since TSR is based on matching lemmatized terms and does not consider their grammatical correctness or the quality of the entire translation. Therefore, a more sophisticated selection strategy is required – one that can maximize the task-specific TSR metric while maintaining overall translation quality.

To achieve this balance, we propose an approach based on Pareto optimality, which we name Pareto-Optimal Decoding. This method identifies the set of candidates that represent the optimal trade-off between general quality, as measured by QE scores, and term accuracy, as measured by TSR. Based on the previously calculated metrics, we pruned the translation candidates to a set of Pareto-optimal hypotheses using the `paretoset`[5] library. From this Pareto set of non-dominated solutions, we selected

[2] https://huggingface.co/Unbabel/XCOMET-XL

[3] https://huggingface.co/ShaomuTan/ReMedy-9B-24

[4] https://huggingface.co/google/metricx-24-hybrid-xl-v2p6

[5] https://github.com/tommyod/paretoset

System	English → German		English → Spanish		English → Russian	
	COMET	**TSR**	**COMET**	**TSR**	**COMET**	**TSR**
Replace	0.8756	0.7882	0.8819	0.9186	0.8570	0.9168
+ new prompt	0.8725	0.8402	0.8764	0.9302	0.8422	0.9304
+ few shot	**0.8862**	0.7846	0.8935	0.9264	0.8764 *	0.9304
+ LoRA + new prompt	0.8834	**0.8528**	0.8980 *	0.9457	0.8851 *	**0.9497**
+ LoRA + new prompt + few shot	0.8853	0.8474	**0.9016**	**0.9477**	**0.8904**	**0.9497**

(a) Translation quality results per language pair.

System	COMET	BLEU	chrF	TSR
Replace	0.8715	40.70	68.34	0.8745
+ new prompt	0.8637	40.59	68.25	0.9003
+ few shot	0.8854 *	43.91 *	70.38 *	0.8805
+ LoRA + new prompt	0.8888 *	45.53 *	70.80 *	**0.9161**
+ LoRA + new prompt + few shot	**0.8924**	**46.24**	**71.34**	0.9149

(b) Macro average results for all language pairs (English → German, English → Spanish and English → Russian).

Table 3: Ablation tests on translation quality for the WMT25 terminology development dataset, comparing individual and combined gains of each method. Results marked with an asterisk (*) are statistically significant compared to the previous method results.

the candidate that maximizes TSR. This final step ensures that our selection directly addresses the primary objective of the Shared Task, while filtering out suboptimal candidates in terms of translation quality as measured by neural metrics.

Figure 1 illustrates our Pareto-Optimal Decoding method for a single source sentence. For the sake of visual clarity, we limited the method to using only two metrics. Each blue circle corresponds to one of the 100 translation candidates, plotted according to the TSR score on the y-axis and `xCOMET` score on the x-axis. The two red × markers highlight the non-dominated solutions.

Interestingly, the hypothesis yielding the highest `xCOMET` score omits the required terminology entirely, resulting in a TSR score of 0.0. This finding emphasizes the limitations of single-metric optimization and the need for a multi-objective approach for translation quality optimization.

4 Results

Initially, we conducted experiments on the WMT25 development dataset. Table 3 shows the improvements gained by using prompt engineering methods and fine-tuning the model using LoRA. Table 3a summarises the results for each language pair using the `COMET`[6] and TSR metrics. Table 3b shows the macro-averaged values for all language pairs, including the `BLEU` (Papineni et al., 2002) and `chrF` (Popović, 2015) metrics.

The results obtained using the replace method were used as our baseline. As Table 3 shows, using a new prompt slightly decreased the general metrics, but improved the term accuracy metric by an average of approximately 2.5 points. Using a few-shot prompt improved results across all general metrics except TSR. Subsequently, applying LoRA further enhanced translation quality, particularly when the model was used with the new prompt and few-shot examples. Although gains for different methods vary between language pairs depending on dataset characteristics, the averaged values indicate that this combined approach yields the best results.

In addition, we performed statistical tests using the Paired Bootstrap Resampling method (Koehn, 2004). We sampled $s = 1000$ times with $n = 0.4 * testset\_length$ segments and p-value $p = 0.05$. Each subsequent processing stage was compared to the previous one. Statistically significant differences are marked with an asterisk (*) in Table 3. The results show that adding few-shot prompting to the baseline solution significantly improved the `COMET` scores for the English → Russian pair, as well as the `COMET`, `BLEU`, and `chrF` scores for the combined dataset for all three language pairs. The second method, which significantly improved the results on `COMET` for the English → Spanish and

[6] https://huggingface.co/Unbabel/wmt22-comet-da

System		chrF ↑	MetricX ↓	ReMedy ↑	xCOMET ↑	TSR ↑
Replace		**68.34**	2.23	0.6203	0.9243	0.898
Pareto	TSR + xCOMET	64.19	1.77	0.6354	**0.9564**	**0.990**
	+ MetricX	64.18	**1.44**	0.6399	0.9524	**0.990**
	+ ReMedy	66.47	1.62	**0.6567**	0.9540	**0.990**
	+ MetricX + ReMedy	65.69	1.53	0.6487	0.9489	**0.990**

Table 4: Comparison of the use of various metrics in Pareto-Optimal Decoding on the WMT25 terminology development dataset. The results are macro-averaged for each language pair.

Mode	xCOMET-QE ↑	ReMedy-QE ↑	MetricX-QE ↓	TSR ↑
English → German				
noterm	0.9907	0.6481	0.6877	0.2413
proper	0.9770	0.6420	1.1579	0.9903
random	0.9798	0.6458	1.0650	0.9913
English → Spanish				
noterm	0.9803	0.6501	1.6306	0.4015
proper	0.9536	0.6472	1.9855	0.9980
random	0.9586	0.6532	1.9721	0.9980
English → Russian				
noterm	0.9844	0.6290	1.3186	0.3113
proper	0.9567	0.6274	1.7454	1.0000
random	0.9624	0.6318	1.5853	0.9980

Table 5: Evaluation of our final submission to the WMT25 Terminology Translation Task. We calculate the TSR for the noterm mode against terminology constraints in the proper mode as suggested by the task organizers.

English → Russian language pairs, as well as on the entire dataset according to the `COMET`, `BLEU`, and `chrF` metrics, was the method using LoRA with a new prompt. The other methods yield improvements when the averaged metric results are compared, but these are not statistically significant.

Table 4 shows the results of the Pareto-Optimal Decoding experiments. Various metrics were tested to evaluate the candidates in this processing stage. The `xCOMET`, `MetricX` and `ReMedy` metrics were employed and combined with the TSR metric. The results demonstrate that Pareto-Optimal Decoding increases the TSR metric value to 0.99 across all considered translation directions, regardless of the selected metrics. Furthermore, we observe an improvement across all translation quality metrics. However, it's important to note that due to metric interference (MINT) phenomenon (Pombal et al., 2025), the evaluation on utility metrics used during Pareto-Optimal Decoding may yield biased results. We hypothesize that our multi-objective approach mitigates the effects of MINT by encouraging the selection of more robust translation candidates. Based on this analysis, we have decided to use `xCOMET` combined with `ReMedy` in the final solution, leaving `MetricX` for fair evaluation.

Experiments showed that the best results were achieved by using a fine-tuned model together with a modified prompt and few-shot examples, as well as by employing the Pareto-Optimal Decoding method along with the `xCOMET` and `ReMedy` metrics. These methods were used to translate the WMT25 test dataset, except for few-shot prompting, which was found not to affect the translation quality when combined with Pareto-Optimal Decoding. The final results are presented in Table 5. For the `proper` and `random` modes, we utilized our full system. For the `noterm` mode, which lacks a terminology list, we used a baseline model with a modified Pareto-Optimal Decoding that relied solely on QE metrics (`xCOMET` and `ReMedy`).

The final results of all the calculated metrics demonstrate that the proposed method performs well across all evaluated metrics. In particular, TSR achieves values above 0.99 in both the `proper` and `random` dataset modes across all considered directions. In English → Russian direction it even reaches 1.0 in the `proper` dataset mode, which means that the entire specified terminology in the dataset was transferred correctly.

References

Adrien Barbaresi. 2021. Simplemma. Archived snapshot of all versions of Simplemma.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *The Twelfth International Conference on Learning Representations*.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kamil Guttmann, Mikołaj Pokrywka, Adrian Charkiewicz, and Artur Nowakowski. 2024. Chasing COMET: Leveraging minimum Bayes risk decoding for self-improving machine translation. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 80–99, Sheffield, UK. European Association for Machine Translation (EAMT).

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation

with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Tommi Nieminen. 2023. OPUS-CAT terminology systems for the WMT23 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 912–918, Singapore. Association for Computational Linguistics.

Artur Nowakowski and Krzysztof Jassem. 2021. Neural machine translation with inflected lexicon. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 282–292, Virtual. Association for Machine Translation in the Americas.

Artur Nowakowski, Gabriela Pałka, Kamil Guttmann, and Mikołaj Pokrywka. 2022. Adam Mickiewicz University at WMT 2022: NER-assisted and quality-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 326–334, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Oumi Community. Oumi: an Open, End-to-end Platform for Building Large Foundation Models.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms.

Ryokan Ri, Toshiaki Nakazawa, and Yoshimasa Tsuruoka. 2021. Modeling target-side inflection in placeholder translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 231–242, Virtual. Association for Machine Translation in the Americas.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling.

Jörg Tiedemann and Lars Nygaard. 2004. The OPUS corpus - parallel and free: `http://logos.uio.no/opus`. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Fine-tuning NMT Models and LLMs for Specialised EN-ES Translation Using Aligned Corpora, Glossaries, and Synthetic Data: MULTITAN at WMT25 Terminology Shared Task

Lichao Zhu Maria Zimina-Poirot Cristian Valdez Stéphane Patin

ALTAE, Université Paris Cité
{lichao.zhu, maria.zimina-poirot, cristian.valdez, stephane.patin}@u-paris.fr

Abstract

This paper presents a hybrid evaluation of terminology-aware English-to-Spanish machine translation systems developed for the WMT25 Terminology Shared Task, specifically targeting the Information Technology (IT) domain. Our objective was to improve terminology accuracy and overall translation quality and highlight the potential of specialised terminology-aware translation models for technical domains. We used different enhancement strategies for both neural machine translation (NMT) systems and large language models (LLMs). These strategies include fine-tuning with synthetic data, the use of in-domain parallel corpora, and hard constraint methods such as placeholder substitution and in-context glossary integration. The results demonstrate distinct lexical and stylistic profiles in the outputs of fine-tuned NMT systems and LLMs, as well as the complementary advantages of different terminology injection methods. Systems behave differently with and without a glossary, as demonstrated by experimental results. The NMT systems seem to be rather limited in adapting to special lexicons and resizing embeddings, which is the opposite of LLMs, which prefer structured instructions. Although our translation systems achieved their highest scores on the *NoTerm, Consistency* metrics, exceeding 81%, demonstrating their ability to produce stable and coherent translations of recurring terms and phrases in unconstrained settings, the precision of the terminology and overall quality of the translation could have been improved by additional training.

1 Introduction

Adaptation of the NMT model to a specific domain (Chu et al., 2017; Chu and Wang, 2018; Saunders, 2021) is a major concern in bilingual neural machine translation. Among the various methods that have been proposed to tackle domain adaptation, two approaches are particularly relevant to the objective of this shared task: i) **Data approach** that involves selecting and filtering existing in-domain parallel segments (Moore and Lewis, 2010; Axelrod et al., 2011), or generating synthetic data. The latter is widely used in back-translation and in enhancing domain-specific data by reformulating or paraphrasing (Sennrich et al., 2016; Edunov et al., 2018); ii) **System approach** which aims to assign weights to segments close to the target domain (Wang et al., 2017). In recent years, research has increasingly focused on frugal domain adaptation strategies, emphasising the efficient use of limited resources and optimising training settings to address low-resource scenarios (Adams et al., 2022; Marashian et al., 2025). With the rapid development of LLMs for translation purposes in recent years, several methods for LLM domain adaptation have emerged, such as prompt engineering and context leaning (Zhang et al., 2023; Pourkamali and Sharifi, 2024; Yamada, 2023), constrained decoding to enforce terminology or special data format (Luca Beurer-Kellner, 2024; Bogoychev and Chen, 2023), Supervised Fine-Tuning (SFT) using reinforcement learning from human feedback (Ouyang et al., 2022), etc. Meanwhile, domain adaptation sheds light on the functioning, strengths, and weaknesses of LLM (Lu et al., 2025a), making these 'black boxes' more interpretable – an important objective of our participation in the shared task. In our contribution, we have chosen to use three strategies to fine-tune both NMT and LLM systems to ensure accurate translation of terms tagged as 'proper_terms' in the dev set:

- An open source NMT system was fine-tuned by employing placeholders for lexical-constrained decoding using additional aligned segments within the domain.
- An artificially augmented training data set

Proceedings of the Tenth Conference on Machine Translation, pages 1284–1291
November 8-9, 2025

was created using a prompt system and used to fine-tune a baseline on a commercial model training server.

- Using aligned segments and a glossary, Low-Rank Adaptation was used to make minor changes to an LLM (Hu et al., 2021).

2 Data processing and augmentation

2.1 Data sources for domain-specific model specialisation

To achieve precise terminology and consistency in fields like finance, IT and legal texts, it is essential to use high-quality aligned corpora and domain-specific glossaries to fine-tune NMT models and LLMs for specialised machine translation. The use of synthetic data generation methods can also help to augment domain-specific corpora, enhancing models' abilities to manage specialised terminology and contexts effectively. Prompt-based generation, retrieval-augmented generation, self-instruction, and reinforcement learning with feedback are some of the approaches available for synthetic data generation (Lu et al., 2025b; Nad et al., 2025). These methods can improve model performance, data diversity, and adaptability to domain-specific requirements by supplementing real training data with synthetic examples generated by LLMs.

2.2 High-quality parallel corpus

To increase the size of the training data and achieve a broader terminological coverage, we used a high-quality parallel corpus in IT and closely related fields: the European Union Intellectual Property Office (EUIPO) Trademark and Design Guidelines in the production, technology and research domains, translated by professional translators from English into Spanish.[1] This resource, created by the European Language Resource Coordination (ELRC), contains 16,439 translation units; 386,472 tokens in English and 424,702 tokens in Spanish. After filtering based on length control and the basic alignment quality of the segments, we obtained 6,359 parallel segments containing 62,481 English tokens and 67,791 Spanish tokens, representing an average of 910 words (60-66 characters) per segment.

[1] https://is.gd/LlPIYr

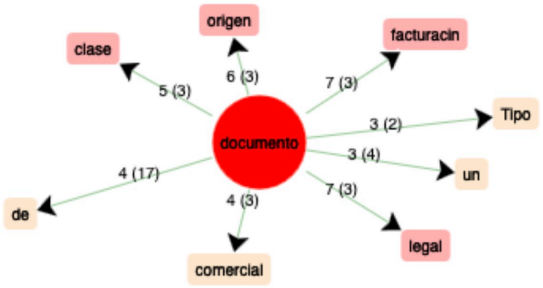


Figure 1: Parallel co-occurrences in the augmented data, measuring the frequency and specificity of lexical attractions.[2]

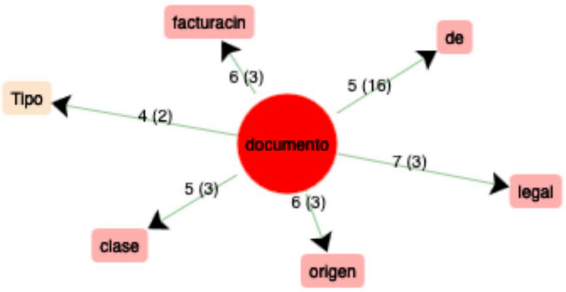


Figure 2: Parallel co-occurrences in the development set of the shared task, measuring the frequency and specificity of lexical attractions.

2.3 In-domain synthetic data

In our experiments, we used the dev data set (enes-dev: 500 aligned segments) augmented via a system of prompts on ChatGPT-4 (first to generate an initial data set in English using task-specific terminology) and on DeepSeek (to translate the generated dataset into Spanish using the WMT 2025 dev set term correspondences). We used free public versions of both platforms. Our objective was not only to consider specific term correspondences, but also to capture linguistic features of the test set to create synthetic data reproducing the dev set patterns. We created a synthetic data set enes-AUG (1,746 aligned English-Spanish sentence pairs, 10,021 tokens in English, 10,822 tokens in Spanish). The generated synthetic data set has some recognisable characteristics in terms of regularly repeated characteristic patterns (for example, in Spanish: *Usa las acciones* ***en tu documento***. *Utilice los Servicios profesionales* ***en su documento***. *Utilice el Soporte SAP* ***en su documento***).

[2] Calculated by iTrameur (textometrics tool): https://itrameur.clillac-arp.univ-paris-diderot.fr

We analysed parallel co-occurrence networks (Zimina and Fleury, 2016) in sentences containing the same terms in the dev set and in the augmented data. It reveals that both data sets share key terminological traits (see Figures 2.3 and 2), supporting the suitability of AI-generated data for our experiment. However, the AI-generated data set lacks the uppercase usage typical of the enes-dev data. For example, the term "Source Document Category" appears in lowercase. Future experiments could include prompt instructions to better replicate capitalisation in terminology.

2.4 Glossary terms and augmented terminology

The enes-dev set contains some rather challenging (or even questionable) annotations of "proper_terms" and "random_terms", for example:

"en": **"Source Document Category"**, "es": **"Tipo de documento de origen"**, "proper_terms": {"source document category": "tipo de documento de origen"}, "random terms": {"Source": "Origen", "Document": "documento", "Category": "tipo"}.

If ***Tipo de documento de origen*** is a term, then its constituents, such as ***documento*** and ***tipo***, are hardly random in specialised discourse, especially if we consider term variation, which involves the use of different forms to express the same or nearly the same concepts within a specialised domain (Daille, 2017). In our work, we considered that such occurrences are part of specialised discourses and their distributional properties (Wingfield and Connell, 2022) are reflected in the augmented data set. We also tried to take into account the fact that different multiword terms possibly contain common lexical items. In this respect, term variation is reflected in the augmented data set. For example, while the term "source document category" is translated by *tipo de documento de origen*, our synthetic data set also contains contexts, where "document type" is translated by *clase de documento*:

en: Save the **source document category**. > es: *Guarda el **tipo de documento de origen***.

en: Save the **document type**. > es: *Guarda la **clase de documento***.

3 Systems

3.1 NMT system

We investigated methods to impose lexical constraints on NMT systems that do not inherently support glossary use. One approach involves fine-tuning a generic NMT model with domain-specific synthetic data generated through LLMs and other generative AI tools. This synthetic data set incorporates targeted terminology to guide system translations, improving sentence level accuracy and ensuring consistent terminology usage. The fine-tuning process leverages data augmentation techniques to integrate specific term correspondences and improve translation coherence within specialised domains.

3.2 LLM

Recent research suggests that LLMs can be effectively tuned for MT using surprisingly small amounts of high-quality parallel data. Fine-tuning with LLMs (such as Llama-2 7B) can deliver strong performance when trained on as few as 32,219 parallel sentences (Lu et al., 2025a). This is in line with the hypothesis of "superficial alignment", which suggests that LLMs have already mastered their translation skills during pre-training, and fine-tuning concentrates on aligning the model with the specific task format.

4 Experimental settings

We used constrained machine translation with hard terminology control (EN-ES) to ensure that English terms were consistently predicted as their Spanish equivalents. We implemented our NMT system with Marian (Junczys-Dowmunt et al., 2018), mainly because it allowed constraint decoding with the additional_special_tokens parameter, and the models embeddings were resized accordingly. In the training data, we replaced English terms and their Spanish equivalents with placeholders: each pair of bilingual terms was replaced by the same placeholder. Consequently, English terms in the input were replaced by placeholders before decoding, and placeholders in the output were replaced by the equivalent Spanish terms after generation. Our system was fine-tuned with Seq2SeqTrainer with the following setting: training epochs: 5, learning rate: 1e-5, training batch size: 8, gradient accumulation steps: 2, max length: 128.

For the LLM model, we used EuroLLM-1.7B-Instruct (Martins et al., 2024) as a causal LM due to its relatively light weight for fine-tuning. For each pair of aligned segments (EN-ES), the system builds a prompt that injects the glossary before instruction as follows:

```
"### Instruction:"
"Translate the following text from English
to Spanish using the provided glossary."
"### Glossary (Information Technology):"
"{glossary_items}"
"### Text in English:"
"{src}"
"### Translation in Spanish:"
"{tgt}"
```

The glossary was created using the English and Spanish lexicons tagged as "proper_term" elements in the dev set. After filtering and deduplication, we obtained 172 unique occurrences in English (not case-sensitive). These occurrences correspond to 221 pairs of English and Spanish terms. Training was set as follows: `training batch size`: 4, `learning rate`: 1e-4, `weight decay`: 0.01, `lr scheduler type`: linear, `warmup steps`: 100, `training epochs`: 1. We used parameter-efficient LoRA (Hu et al., 2021) adapters for light fine-tuning, using 8-bit loading to reduce GPU memory usage: `task type`: CAUSAL LM, `r`: 16, `lora alpha`: 32, `lora dropout`: 0.1. For both the Marian and EuroLLM fine-tuning processes, we used a small number of epochs and a small batch size to reduce training time and prevent over-fitting. The main purpose was to evaluate the efficiency of our training pipelines.

We also used an advanced commercial MT platform SYSTRAN Model Studio Lite[3] to fine-tune a generic EN-ES baseline model. NMT was used to ensure accurate sentence-level translations. GenAI, including LLMs, was used to develop a synthetic data set that contains domain-specific terms to fine-tune the initial system and implement specific term correspondences, improving coherence regarding terminology specifications (see Section 2.3 on data augmentation techniques).

[3] https://modelstudio-lite.systran.net

5 First results

5.1 Quantitative metrics

We investigated lexical characteristics and vocabulary usage in six translations produced by different models (see Table 1) and compared them using quantitative methods, such as `correspondence analysis`, `vocabulary growth assessment`, and `characteristic elements computation` (Lebart et al., 1998). The investigation revealed nuanced differences in the translations and highlighted the impact of glossary inclusion and placeholder handling on translation quality and style.

A corpus aggregated from six different translations contains a total of 35,250 tokens. On average, each translated text comprises approximately 5,875 tokens, with two notable exceptions: **EuroLLM_SEG_Glossary** (6,192 tokens) and **MTwithplaceholder** (5,400 tokens).

5.1.1 Vocabulary growth

The vocabulary growth curves, which reflect the natural increase in distinct words encountered, show that **EuroLLM_SEG_Glossary** consistently has the largest vocabulary size (see Figure 3). This suggests that it has the most diverse vocabulary of all the models tested. **Systran_SEG_AUG** exhibits somewhat lower vocabulary growth, very close to the reference translation, while **Marian_SEG** shows the slowest increase, suggesting a smaller or more limited vocabulary compared to the others.

5.1.2 Characteristic elements

The results of `characteristic elements` computation reveal the use of English lexical units in **MTwithplaceholder**, which do not correspond to common Spanish borrowings, highlighting the models lexical weaknesses. **EuroLLM_SEG_Glossary** distinctly exhibits a preference for informal address forms, favouring pronouns and verbs such as *puedes*, *selecciona*, and *tus* over their formal counterparts like *pueda* or *Seleccione*. It also underuses the term *plantilla*, substituting it with *modelo* or omitting it entirely, indicating stylistic or glossary-driven variations.

5.1.3 Correspondence analysis

`Correspondence analysis` provided a multidimensional visualisation of the lexical relationships

[4] Generated with iTrameur (textometrics tool): https://itrameur.clillac-arp.univ-paris-diderot.fr

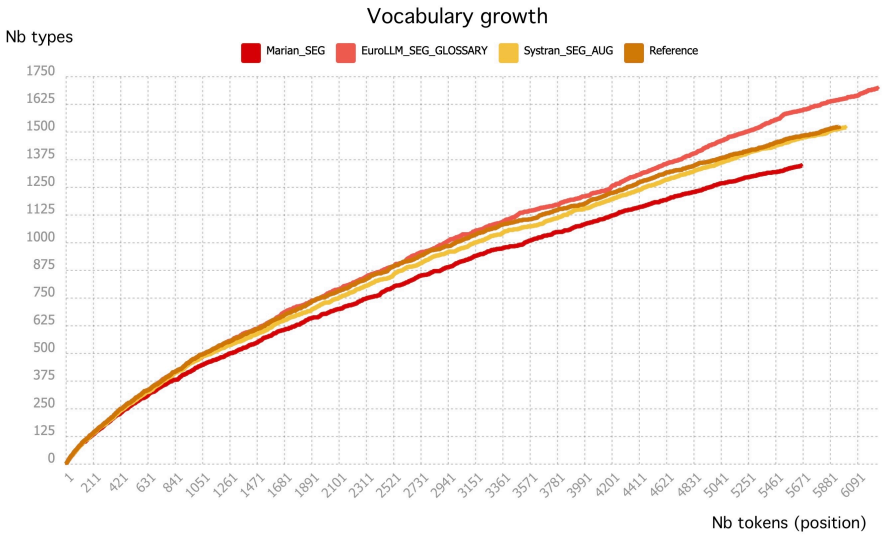


Figure 3: Comparison of vocabulary growth curves among submitted translations and the reference.[4]

across the models (see Figure 4). The first factorial plane, which accounts for 49.9% of the total lexical variation, highlights the following oppositions:

- **(Pure) MT translations vs. (human) reference**: As expected, the results of the correspondence analysis highlight the differences between the reference translation and the translations produced by the submitted MT systems.
- **NMT vs. LLM**: Submitted MT translations are positioned along a continuum that differentiates between NMT and LLM characteristics. The **Systran_SEG_AUG** system, which relies on LLM-generated data, exhibits hybrid traits.

The second factorial plane (see Figure 4), which accounts for 26.7% of the total lexical variation, further confirms oppositions and proximities observed in the first plane, mapping a closer proximity between **Systran_SEG_AUG** and the reference translation compared to the other submitted translations:

- **Systran_SEG_AUG** shows some characteristics that approximate the **reference** translation.

5.1.4 Performance scores

Following this analysis, we conducted an evaluation using the reference translations provided by the organisers. Table 2 compares the performance of pre-trained or generic models (Marian MT, EuroLLM, and Systran *Generic*) with that of fine-tuned models. Our analysis shows that although constrained decoding helps models better translate "proper terms", it reduces the overall translation performance of MT systems. This effect is particularly evident in Marian MT: although the fine-tuned model translated more proper terms and produced more formal equivalences, its overall translation quality decreases, with the COMET-DA score dropping from 0.86 (pre-trained) to 0.82 (fine-tuned). However, constrained decoding improves the term-matching precision of EuroLLM without heavily degrading its overall translation quality. For Systran, the score difference between the pre-trained and fine-tuned models is minimal.

5.2 Linguistic analysis

Evaluating fluency is difficult since most segments are decontextualised noun phrases, but there are examples where fluency differences between models are noticeable (see Table 3).

Systran_SEG_AUG tends to produce translations that are quite literal but accurately preserve the source content. The system generally yields the best results in terms of information transfer. The **Marian_SEG** model sometimes omits segments. When this happens, there are negative effects on orthographic correctness, such as missing capital letters at the beginning of sentences.

[5]Generated with **Lexico 5**, lexicometrics software: https://lexi-co.com/L5Presentation.html

[6]*: Submitted version.

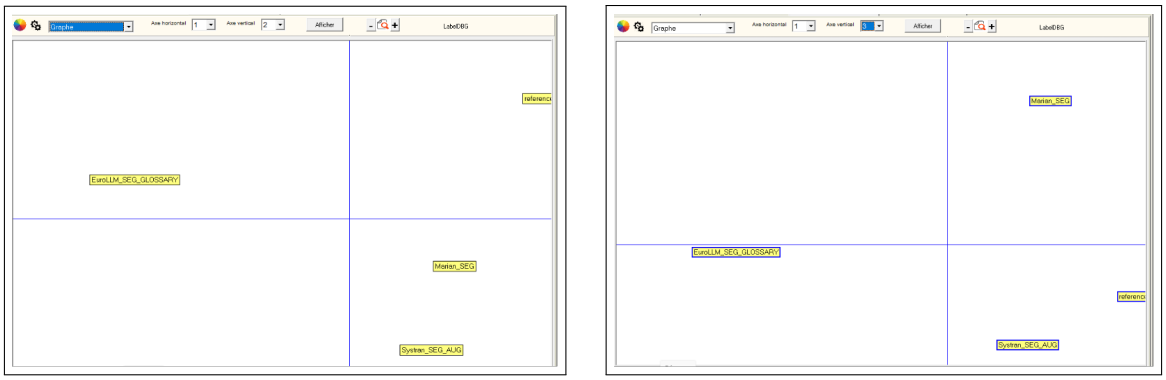


Figure 4: Factorial correspondence analysis of submitted translations. The first factorial plane, formed by axes 1 and 2, accounts for 49.9% of the variation. The second one, formed by axes 1 and 3, accounts for 26.7%.[5]

System	Token	Form	Hapax	Fmax
ES_SystranIT	6,014	1,519	834	653
ES_Systrangeneric	5,976	1,504	849	677
EuroLLM_SEG_GLOSSARY*[6]	**6,192**	**1,670**	**990**	**690**
MTwithplaceholder	5,400	1,592	973	440
Marian_SEG*	5,664	1,351	762	658
Systran_SEG_AUG*	6,004	1,524	866	684
Total	35,250	2,871	873	3,802

Table 1: Quantitative characteristics of the translations generated by each of the models.

EuroLLM_SEG_GLOSSARY model occasionally leaves some English segments untranslated. This occurs rarely but more often than with other models. Some untranslated segments result from poor segmentation in the source text. There is no consistent correlation between the omission or retention of expected terms across models. For example, sometimes expected terms are omitted, sometimes other segments are omitted leaving only the expected term, and in some rare cases the source text remains unchanged.

In cases where **EuroLLM_SEG_GLOSSARY** leaves English segments untranslated (e.g., "On the To Be Billed"), it appears to stem from improper segmentation. For example, the segment "On the To Be Billed" that is left in the translation generated by **EuroLLM_SEG_GLOSSARY** is the result of poor segmentation of the original version. If the decision was to leave the tab name in English, it should have been rendered as "*en la pestaña To Be Billed*". **EuroLLM_SEG_GLOSSARY** sometimes introduces lexical creations or barbarisms, which can make understanding difficult.

In summary, **Systran_SEG_AUG** performs best in preserving source information. **Marian_SEG** suffers from segment omission causing orthographic errors. **EuroLLM_SEG_GLOSSARY** tends to leave English untranslated more often and sometimes produces confusing lexical forms.

5.3 Interpretative insights

This combined quantitative and qualitative analysis shows that translation models vary in more than just lexical richness; they also differ in terms of stylistic choices and lexical specificity. There are significant differences in how models handle formality, lexical borrowing, and terminology consistency, reflecting variations in MT architecture and preprocessing (e.g. glossaries and placeholders). Correspondence analysis is a useful tool for visualising these differences, as it reduces complex lexical data into interpretable axes of variation. This enables more informed evaluations of MT output.

6 Conclusion and perspectives

Our contribution to the shared task reveals the strengths and challenges of NMT systems and LLMs in translation tasks. Although NMT systems perform well with small amounts of high-quality domain-specific training data, their performance can deteriorate under constrained decoding conditions. In contrast, LLMs can benefit from structured guidance, such as glossaries and clear instructions, which enhances their translation quality. These findings emphasise the complementary

System	Total terms	Translated terms	Ratio	BLEU	chrF	COMET DA
EuroLLM_SEG_GLOSSARY	538	204	37.9%	38.6	69.1	0.83
EuroLLM_1.6B_Pre-trained	538	199	37.0%	36.1	53.4	0.84
Marian_SEG	538	**254**	**47.2%**	48.1	73.2	0.82
Marian_Pre-trained	538	196	36.4%	45.0	58.1	0.86
Systran_SEG_AUG	538	222	41.3%	50.7	**75.7**	**0.88**
Systran_Generic	538	221	41.1%	**51.2**	73.2	0.88

Table 2: Term coverage and evaluation scores

ID	Source text	Systran_SEG_AUG	Marian_SEG	EuroLLM_SEG_GLOSSARY	Term
1	On the To Be Billed Tab, select one or more items as required and choose Write Off.	En la ficha Para facturar, seleccione uno o más elementos según sea necesario y elija Cancelar.	En la pestaña Para ser facturado seleccione uno o más elementos como sea necesario y seleccione Escribir apagado.	En la pestaña On the To Be Billed, seleccione uno o más elementos según sea necesario y elija Deshacer.	Write off<>ignorar
2	On the To Be Billed Tab, select one or more items as required and choose Restrict Date.	En la ficha Para facturar, seleccione uno o más elementos según sea necesario y Restringir fecha.	En la pestaña Para ser facturado seleccione uno o más elementos como sea necesario y seleccione Limitar fecha.	En la pestaña On the To Be Billed, seleccione uno o más elementos según sea necesario y Restrict Date.	Tab<>pestaña

Table 3: Translation comparisons

nature of the two approaches, suggesting that combining NMT's data-driven learning with LLMs' flexible, instruction-driven capabilities could result in more robust and effective translation solutions.

We also identified some limitations through our experimentation. None of our open-source systems translated all the segments from English to Spanish, with some segments always remaining untranslated. This is probably due to gaps in the lexicons of the training and test data or to forced decoding effects that prioritise and select the source token over the target token. In summary, a trade-off must be made between model complexity, performance, and training cost. While a 16B-parameter model could potentially avoid missing any translations, it is much more challenging to fine-tune such a model due to its substantial number of parameters, size, and the associated high training costs in terms of GPU memory, time, and environmental impact.

7 Acknowledgements

This research was funded by the 2024 **MULTITAN-GML** Research Equipment Grant (COPES-2024-12, *Fonds d'intervention Recherche, Université Paris Cité*) and used **PNS-UP** scientific platform[7].

References

Virginia Adams, Sandeep Subramanian, Mike Chrzanowski, Oleksii Hrinchuk, and Oleksii Kuchaiev. 2022. Finding the right recipe for low resource domain adaptation in neural machine translation. *Preprint*, arXiv:2206.01137.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An empirical comparison of domain adaptation methods for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Chenhui Chu and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. In

[7] https://plateformes.u-paris.fr

Proceedings of the 27th International Conference on Computational Linguistics, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Béatrice Daille. 2017. *Term Variation in Specialised Corpora: Characterisation, Automatic Discovery and Applications*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Ludovic Lebart, Salem André, and Berry Lisette. 1998. *Exploring Textual Data*. Academic Kluwer Publisher, Dordrecht, Boston, London.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. 2025a. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 11(84).

Yingzhou Lu, Lulu Chen, Yuanyuan Zhang, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2025b. Machine learning for synthetic data generation: A review. *Preprint*, arXiv:2302.04062.

Martin Vechev Luca Beurer-Kellner, Marc Fischer. 2024. Guiding llms the right way: Fast, non-invasive constrained decoding. *arXiv*.

Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. From priest to doctor: Domain adaptation for low-resource neural machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 220–224.

Mihai Nad, Laura Dioan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615134633.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *Preprint*, arXiv:2401.08429.

Danielle Saunders. 2021. *Domain Adaptation for Neural Machine Translation*. Ph.d. thesis, University of Cambridge, Cambridge, United Kingdom. Available online.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. 2017. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488.

Cai Wingfield and Louise Connell. 2022. Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, 37(10):1220–1270.

Masaru Yamada. 2023. Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 195–204, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

Maria Zimina and Serge Fleury. 2016. Perspectives de larchitecture trame/cadre pour les alignements multilingues. *Nouvelles perspectives en sciences sociales*, 11(1):325–353.

Contextual Selection of Pseudo-terminology Constraints for Terminology-aware Neural Machine Translation in the IT Domain

Benjamin Pong
University of Washington
Seattle WA, USA
{benpong}@uw.edu

Abstract

This system paper describes the development of a Neural Machine Translation system that is adapted to the Information Technology (IT) domain, and is able to translate specialized IT-related terminologies. Despite the popularity of incorporating terminology constraints at training time to develop terminology-aware Neural Machine Translation engines, one of the main issues is: In the absence of terminology references for training, and with the proliferation of source-target alignments, how does one select word alignments as pseudo-terminology constraints? The system in this work uses the encoder's final hidden states as proxies for terminologies, and selects word alignments with the highest norm as pseudo-terminology constraints for inline annotation at run-time. It compares this context-based approach against a conventional statistical approach, where terminology-constraints are selected based on a low-frequency threshold. The systems were evaluated for general translation quality and Terminology Success Rates, with results that validate the effectiveness of the contextual approach.

1 Introduction

This paper describes UW-BENMT's submission[1] to WMT2025 Terminology Translation Shared Task (Semenov et al., 2025). The aim of this edition of the shared task is to evaluate how well machine translation systems can handle specialized terms in specific domains where terminology accuracy and consistency are critical. Although general machine translation systems have improved significantly, specialized terminology remains a challenge (Semenov et al., 2023; Alam et al., 2021), and this task evaluates the effectiveness of integrating terminology dictionaries into machine translation engines in the Information Technology (IT) domain. This is a highly practical task as machine translation engines that have been adapted to this domain can potentially assist in the international communication of technical APIs, manuals and DevOp guides across IT teams that operate in multilingual environments. This task involves segment-level terminology translation for three language pairs: English $\rightarrow$ {German, Spanish, Russian}. The provided data consists of 500 parallel sentences per language pairs, with reference terminology dictionaries.

2 Related Work

There are two major approaches to terminology translation, with the first being lexical constrained decoding where the target terminologies entries are forced to match the source side lexical terms as decoding-time constraints (Chatterjee et al., 2017). However, one major shortcoming of this strategy is the computational overload as the number of terminology constraints increase. To address this issue, Dinu et al. (2019) pioneered a methodology to train neural machine translations engines to recognize terminology constraints. The focus of their approach is to annotate the data with source-target terms inline as soft constraints (i.e., a form of data augmentation). The success of this approach is reflected in popularity of system submissions that adopted it at the WMT2021 and WMT2023 Terminology Translation Shared Tasks (Ailem et al., 2021; Nieminen, 2023; Bogoychev and Chen, 2023; Park et al., 2023). Different alternatives to implementing this approach have been proposed, such as masking out the source-side terms (Ailem et al., 2021; Liu et al., 2023), which have been argued to surpass simple inline term annotations.

As pointed out by Bogoychev and Chen (2023), one of the main challenges of terminology transla-

[1]Repository for system development can be found at https://github.com/Benjamin-Pong/Terminology-Neural-Machine-Translation-IT-domain.git

Proceedings of the Tenth Conference on Machine Translation, pages 1292–1301
November 8-9, 2025

tion is that terminology dictionaries are not readily available for training, and their proposal differs from Dinu et al. (2019) in that they devised a way to automatically create terminology constraints for existing training corpora. The development of pseudo-terminology constraints is a challenge in itself because it is difficult to select word alignments that are most representative of 'technical terminologies'. While Dinu et al. (2019) and Bogoychev and Chen (2023) used randomly selected terms, other systems construed domain-specific terms as low frequency lexical items (Semenov et al., 2023; Park et al., 2023), a common approach that is practiced in works beyond the shared task (Koehn and Knowles, 2017; Yuan et al., 2018; Bowker, 2021).

The main contribution of this system paper is to shed light on the context-driven approach to selecting pseudo-terminology constraints for IT-adapted NMT training, and present this as an alternative to the frequency-based approach. It focuses on improving the psuedo-terminology creation process in the absence of training terminology dictionaries by comparing two methods; frequency-based, context-based.

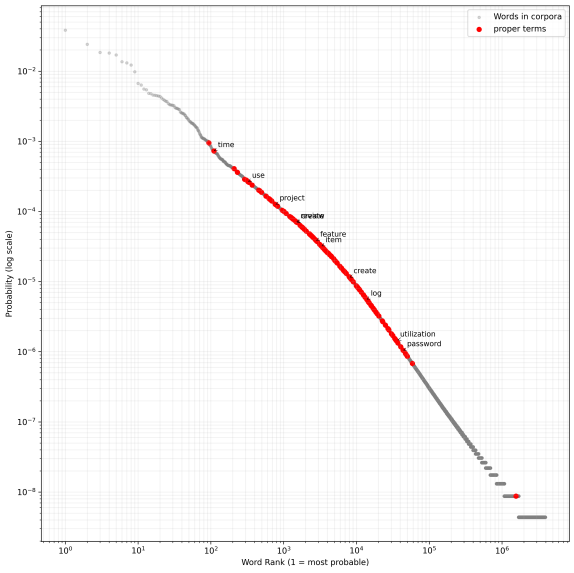


Figure 1: Rank-frequency plot showing the frequency distribution of proper terms against all words in a corpora

3 Selection of Pseudo-terminology Constraints

3.1 Proxies for Terminology Selection: Frequency vs Context

There are two main reasons why frequency may not be the best proxy for terminology selection. The first reason is motivated by the statistical frequencies of proper terms against the tokens in publicly available corpora. Consider the Rank-Frequency Distribution in Figure 1, where the frequencies of the provided data's English terminologies were compared against the frequencies of all English tokens from the Europarl and WikiMatrix data[2]. The frequencies were rescaled to a value between 0 to 1. Based on the Rank-Frequency distribution of terminologies, terminologies lie in mid-rank, contrary to conventional assumptions that terminologies are low-frequencies.

The second reason lies in the fact that lexical terms selected as proper terms are not treated as proper terms in all contexts. A vast majority of the proper terms in the development set are also treated as random terms and do not have any domain-specific meanings. Consider the way 'create', a mid-ranked frequency word based on the rank-frequency distribution, is used in following sentences taken from the provided data: (1) *With the Story Builder, you can create stories to visualize information with charts and table* and (2) *Activate or Create your Data Provider Profile.*

"Create" is treated as a specialized IT-terminology in the first sentence, but not in the second sentence. This ambiguity is difficult to resolve but one could speculate that "create" in the first instance is used in the context of a software application assisting with the creation, which renders it a specialized IT-terminology. Since it is challenging to define deterministic semantic rules for domain-specific terminologies, and neither does a frequency-based approach suffice since 'create' is not exactly a low-frequency word, an approach that is agnostic to these two conditions should be explored.

3.2 Terminology Constraint Filtering by Norm of Top Most Hidden States

Inspired by Wu et al. (2024) and Schakel and Wilson (2015) who established a theoretical and empirical connection between important-words and context for general machine translation, this paper adapts their methodology to terminology-aware machine translation. More specifically, the methodology is applied at the pseudo-terminology constraints creation stage of the end-to-end NMT pipeline, where source-target alignments are filtered based on how "important" they are in seman-

[2]There are a total of 4 million unique English words in both of these corpora.

Language Pair	Dataset	Sentences	Out-domain Training	Out-domain For CED filtering
en-es	WikiMatrix	6,452,177	30,000	8,303,595
	Europarl	1,881,418		
	Total	**8,333,595**		
en-de	WikiMatrix	6,227,188	30,000	8,025,709
	Europarl	1,828,521		
	Total	**8,055,709**		
en-ru	WikiMatrix	5,203,872	30,000	10, 551, 783
	ParaCrawl v9	5,377,911		
	Total	**10,581,783**		

Table 1: This table shows the dataset sources and statistics for each language pair. It also shows the amount of data that was used for CED filtering, from which approximately 2.2 million samples were selected for training purposes.

tic meaning. The norm of the encoder's final hidden states represents contextually important information within lexical items (Wu et al., 2024).

While the importance of context in resolving ambiguities in machine translation (Maruf et al., 2019; Post and Junczys-Dowmunt, 2024) is not entirely new, its effects on terminology translation has yet to be explored, which is the focus of this system paper. Given a set of source-target alignments for a sentence-pair, and given that the norm of the encoder's final hidden states is computed for each source word, the word-alignment pair(s) whose source-word has the highest norm will be chosen as a pseudo-terminology constraint for augmentation.

4 Neural Machine Translation Training

4.1 Training Data: IT-domain Parallel Data Selection using Cross-Entropy Difference

Due to data sparsity in the IT-domain, it is necessary to procure sufficient quality data for system developments. To do so, Cross Entropy Difference (CED) (Moore and Lewis, 2010) was applied to select IT-specific data from a larger corpus of out-of-domain or general domain content.

For CED to be effective, the assumption is that there is already some available in-domain data. Inspired by Moslem et al. (2023), LLM was used to generate synthetic monolingual (English) in-domain parallel data using the gold terminologies provided by the organizers. The data provided contains 500 parallel sentences per language pair, and each sentence-pair comes with a set of terminology mappings. Using only the source-side terminologies as inputs into Aya-expanse-8b (Dang et al., 2024) with temperature=0 and top-p sampling of 1, synthetic monolingual data was generated with the following prompt: `Please use the terms '{terms}' to generate {number_of_sents} full sentences in {source_language}-{target_language} whose content is related to information technology.`

Note that eventhough parallel data was generated, for the purposes of the task at hand, the English source sentences were used for subsequent indomain and outdomain language modeling. A total of 30K IT-related synthetic sentences were generated and treated as in-domain data. The average sentence length is 50-75 tokens.

Table 1 shows the sets of publicly-available parallel corpora that were considered for each language pair. These serve as the general-domain content. For each set of corpora, 30K English sentences were randomly sampled to train an out-domain 4-gram language model (with Laplace Smoothing). To ensure that the in-domain and out-domain models are comparable, apart from controlling for the size of training data, the sentences selected for out-domain model training also had an average of 50-75 tokens.

The remaining sentences were scored according to the cross-entropy difference between the in-domain and out-domain language models, and ranked in non-increasing order since lower entropy signifies a closer match to the in-domain data[3]. Throughout this process, several preprocessing strategies were applied to reduce noise in the publicly available data. FastText (Joulin et al., 2016) with a threshold of 0.9 was used to remove sentence-pairs where either the source or target sentence was not the desired language. Sentence-

[3]Discussion of CED is beyond the scope of this paper. I encourage the reader to refer to Moore and Lewis (2010) for details.

pairs that contained only punctuations and numbers were also removed. The top 20% was chosen as in-domain training data, and the bottom 20% as out-domain data, which amounts to 2.2 million parallel sentences per language pair. Additional checks were performed to ensure that parallel sentences chosen for training did not overlap with the provided data so that the provided data can be held-out for evaluation.

The overall effectiveness of CED for data selection was validated on downstream general translation quality, and hence the main approach to data selection for subsequent experiments for terminology translation. See Section 6 for details.

4.2 Experiments

The baseline uses the training data as they are, without any lexical constraints being augmented inline at training time. Since the chief focus of this paper is to compare the effects of using (low)frequency and encoder-based contextual scoring for pseudo-terminology selection, two more experimental configurations were designed. The first being training data where only low frequency words were selected as pseudo-terminologies for augmentation (i.e FreqTerm). The second configuration consists of data where only words with the highest norm computed from the final layer of the encoder's hidden states (i.e ContexTerm). For these two configurations where lexical constraints are augmented inline, similar to Dinu et al. (2019), only 10% of the training data is augmented, amounting to around 200K randomly selected sentence-pairs. The special token <*src*> was used to mark the start of the source word and <*tgt*> was used to mark the start of the corresponding target word.

The first part of creating the pseudo-terminology involved extracting word alignments between the source and target languages by using a state-of-the-art neural word aligner, Awesome Align (Dou and Neubig, 2021). Alignments between stopwords were removed. Given that awesome align does not account for multi-word alignment, further lightweight processing was applied to merge consecutive source words that were mapped to the same target word to produce multiword source-target word alignments.

To address the proliferation of word-alignments, the next step is to implement the frequency and contextual approaches for selecting word-alignments to be used as pseudo-terminology constraints.

The former was straightforwardly implemented by first computing a frequency distribution (normalized to values between 0 and 1) for the source-side words using the source-side's training data. Only source-target aligned pairs whose source words' probabilities were at most 10^{-5}, were selected. This low-frequency threshold is not arbitrary but instead, motivated by the aforementioned frequency distribution of terminologies (See Section 3.1 and Figure 1). For a multiword source word, if a subword met the threshold, the entire multiword source word was chosen as a suitable candidate for term annotation. Furthermore, since each sentence may have multiple candidate word-aligned pairs that meet this threshold, randomly chosen aligned word-pairs were chosen for augmentation per sentence. Note that the number of word aligned pairs chosen for augmentation can be adjusted by the engineer.

As for the context-scoring configuration, a pre-trained encoder-decoder NMT model (Ng et al., 2019) was used to extract the final layer's hidden representation per token. Only the source-side (English) needs to be encoded for all language pairs. For each multiword source word, max-pooling is applied to compute a unified final hidden representation, followed by a computation of the norm of this hidden representation. These words were ranked in non-increasing order, with the source word with the highest norm being selected, and consequently, its corresponding source-target alignment, was selected as a terminology constraint.

To ensure a fair comparison between FreqTerm and ContexTerm, two factors were considered; first, the number of terminology constraints selected per sentence for augmentation, and the number of sentences that contain the constraints. The latter was standardized across both experiments by choosing only 10% of the data to be augmented. The former was enforced by enforcing identical thresholds on hyperparameter *k*; k-randomly selected low frequency word(s) and the top-k highest norm(s).

The training data per language pair were tokenized using Moses (Koehn et al., 2007). Byte Pair Encoding (BPE) for subword segmentations (Sennrich et al., 2016) were independently applied to both the source and target languages with 32000 merge operations. 500,000 sentences from the training data were selected to generate BPE codes.

4.3 Model Architecture

All systems were Transformer-based networks, trained (Vaswani et al., 2023) using *fairseq* (Ott et al., 2019). Training configurations were specif-

ically optimized for each language pair. See Appendix A for the hyperparameters. En-es and en-ru were trained for minimum of 35 epochs and a maximum of 50 epochs with early stopping, while en-de was trained for a minimum of 50 epochs and a maximum of 100 epochs with early stopping. Uniformly across all models, the last 7 checkpoints were averaged and used for decoding, with a beam size of 5.

5 Evaluation

Systems were evaluated on the provided data, which was not used for training purposes, and have been verified to not overlap with the selected training data. Although examples were taken from the data to illustrate the point in Section 3.1, note that the approach in this paper does not rely on defined patterns or rules from the data, resulting in minimal risk of evaluation bias. Furthermore, the provided data is the only source of gold references for quality evaluation. With these careful considerations, the provided data was safely treated as held-out test data.

5.1 Evaluation Metrics

The translations provided by the trained models were scored against the gold references using BLEU (Papineni et al., 2002), Chrff++ (Popović, 2017) and COMET-DA (Rei et al., 2020) to assess general translation quality. Terminology Success Rate (TSR) (Semenov et al., 2023) was also used to assess the degree of occurrences of the term translations, with and without lemmatization to account for morphological complexities across the different languages, and also fuzzysearch with a threshold of 90% to account for orthographic deviations. Another reason for selecting this mode of evaluation instead of exact matching (Alam et al., 2021) is to capture adequate term translations, and also to consider the effect of syntactic contexts on the morphological shape of the terms.

5.2 Modes of Evaluation

The NMT systems were evaluated under three modes of incorporating terminology constraints at inference time, with the first mode being no terminologies (i.e., *no term*) where the source sentence is freely translated into the target language. The second mode requires the *proper* terminologies to be incorporated, and the third mode incorporates *random* source-target word alignments.

Results from NMT engines are compared between these three modes. The purpose of incorporating random source-target word alignments is to ensure that any improvements over the *no term* setting brought about by incorporating proper terminologies are not by-products of superior general translation quality (Semenov et al., 2023).

6 Results

Table 2 shows the results of all three systems (baseline, freqTerm and ContexTerm) for all language pairs across the three terminology modes. Table 3 shows the validation results for CED-selected data without any terminology incorporation for training or inference.

6.1 Translation quality

From the results in Table 2, the highest scores for translation quality metrics tend to skew toward the ContexTerm system, with the exception of en-es language pair, where FreqTerm achieved the highest BLEU and ChrF++ scores. Notably, none of the baselines achieved the highest scores.

With regards to how the terminology modes affect overall translation quality, based on the results, the highest scores tend to skew towards systems with *proper* terminology settings. However, there are several exceptions. For instance, the *no term* setting for ContexTerm has the highest BLEU score for en-es. In a similar vein, among the baseline systems, incorporating random terms or proper terms at inference time tends to result in a drop in translation quality.

6.2 Terminology Success Rate

The incorporation of soft constraints at training time increased the Terminology Success Rate (TSR) by a huge margin compared to the baselines. This is demonstrated by the fact that for all language pairs, the *proper* terminology setting achieved the highest TSR compared to the *random* and *no term*. The latter two settings tend to perform comparably low. This pattern is observed for both FreqTerm and ContexTerm, regardless of whether lemmatization was employed. These results are expected and aligns with the findings of previous works mentioned in Section 2.

Crucially, using *proper* terminology setting as a basis for comparison between FreqTerm system and ContexTerm system, the latter has the highest TSR.

Language	System	Modes	BLEU	ChrF++	COMET	TSR^{-L}	TSR^{+L}
en → es							
	Baseline	No Term	27.79	56.60	0.731	0.219	0.233
		Random	18.18	47.93	0.474	0.189	0.198
		Proper	22.35	50.78	0.482	0.421	0.423
	FreqTerm	No Term	27.42	55.79	0.732	0.222	0.232
		Random	26.08	55.20	0.720	0.227	0.238
		Proper	**29.83**	**57.94**	0.713	0.512	0.525
	ContexTerm	No Term	27.62	56.13	0.700	0.233	0.250
		Random	26.08	56.22	0.726	0.229	0.240
		Proper	27.75	57.84	**0.734**	**0.663**	**0.678**
en → ru							
	Baseline	No Term	21.02	54.53	**0.830**	0.251	0.329
		Random	19.98	52.51	0.812	0.232	0.309
		Proper	20.37	53.28	0.810	0.237	0.326
	FreqTerm	No Term	23.46	53.20	0.807	0.248	0.327
		Random	22.98	53.56	0.810	0.249	0.337
		Proper	22.63	53.91	0.811	0.489	0.571
	ContexTerm	No Term	**24.39**	53.34	0.814	0.248	0.327
		Random	22.66	53.93	0.812	0.248	0.325
		Proper	23.22	**55.34**	0.820	**0.574**	**0.654**
en → de							
	Baseline	No Term	13.91	41.96	0.656	0.151	0.142
		Random	13.21	40.0	0.41	0.150	0.141
		Proper	14.09	42.00	0.661	0.151	0.147
	FreqTerm	No Term	13.62	41.48	0.665	0.153	0.147
		Random	13.00	42.67	0.653	0.154	0.145
		Proper	13.56	44.05	**0.672**	0.654	0.645
	ContexTerm	No Term	13.10	41.04	0.66	0.152	0.150
		Random	14.43	45.23	0.670	0.177	0.171
		Proper	**14.62**	**45.98**	0.641	**0.774**	**0.755**

Table 2: Evaluation of systems by language direction, augmentation approaches and terminology modes across BLEU, ChrF++, COMET and Terminology Success Rates(TSR). TSR^{-L} refers to non-lemmatized setting while TSR^{+L} refers to lemmatized setting. Best performing systems and with their corresponding terminology settings are bolded.

Language	Domain	BLEU	ChrF++	COMET	TSR^{-L}	TSR^{+L}
en → es						
	in	27.79	56.60	0.731	0.219	0.233
	out	24.92	52.464	0.767	0.200	0.212
en → ru						
	in	21.02	54.53	0.830	0.251	0.329
	out	20.69	53.57	0.826	0.248	0.317
en → de						
	in	13.91	41.96	0.656	0.151	0.142
	out	15.50	45.68	0.69	0.172	0.194

Table 3: This table shows the evaluation of 'IT-domain' CED-selected data vs out-domain data without terminology constraints incorporated for training or inference.

With respect to the effect of lemmatization, TSR tends to be better with lemmatization for en-es and en-ru. Such improvements are consistent across all terminology modes as well, and across both systems. However, the degree of improvement is language-dependent. For instance, in the context of en-ru (*proper*), the *lemmatized* setting surpassed the *non-lemmatized* setting by at least 0.1 points. This is in stark contrast with en-es (*proper*) where the *lemmatized* setting only surpassed the *non-lemmatized* setting by 0.02 points. Interestingly, for en-de, we observe an opposite effect, where the *non-lemmatized* setting surpasses the *lemmatized* setting by a small margin. This is consistent across both systems, and across terminology modes. See Section 7 for possible explanations and implications of evaluating the adequacy of terminology translation with or without lemmatization.

6.3 Data Selection by Cross-Entropy Difference

Table 3 shows that CED-selected data scored higher in general domain translation tasks. In addition, even without terminology incorporation, using in-domain data achieved relatively higher TSR. However, German is an exception, which will be discussed in Section 7.

7 Discussion

7.1 Contextual Selection of Terminology Constraints

The comparisons between FreqTerm and ContexTerm clearly show the effectiveness of using contextually-based scoring to distill quality word-alignments to be used as terminology constraints or inline term annotations. ContexTerm surpasses FreqTerm by 0.08-0.15 points. This suggests that terminologies are construed in terms of the way they are being used in an utterance (i.e the 'create' example from Section 3), and not completely dependent on frequencies.

7.2 Effect of Contextually-selected Terminology Constraints on Translation Quality

The use of contextually-selected constraints seem to have a trickle-down effect on translation quality as well. As noted previously, systems with terminology constraints at training time tend to have higher translation quality. While this is expected (Dinu et al., 2019), the fact that BLEU , ChrF++ and COMET scores tend to be higher in ContexTerm compared to FreqTerm across all terminology modes suggest that these translation engines are learning domain-specific lexical alignments, which reduces variability, and thereby improving the translation outputs.

7.3 Terminology Success Rate: Lemmatization

The asymmetrical results between en-es and en-ru against en-de with regards to lemmatization could be attributed to the morphological properties of the languages. German, compared to Spanish has been argued to have more complex inflectional paradigms, and that simplifying them improves word alignment (Axelrod et al., 2008). This suggests that lemmatization should correlate with improved scores, but this is not the case. Perhaps the lemmatization oversimplified certain lexical items, resulting in noise. This is an interesting case and should be left for future work through error analysis.

7.4 Data Selection by Cross-Entropy Difference

The fact that CED-selected data showed marginally higher general translation and TSR for en-es and en-ru may indeed suggest that this is an effective approach to source quality parallel data for a less resourced domain. Furthermore, the higher TSR scores suggest that training on data closer to the actual domain can boost terminologies, without the incorporation of constraints at train time. However, the converse results for en-de, although disappointing, may be due to actual noise in the data. This investigation will be left for future work.

8 Conclusion

This paper addresses a challenge in terminology translation engines that incorporate term annotations at run-time: with the proliferation of source-target word alignments, how does one select the optimal subset of alignments as pseudo-terminology constraints? Results from this work show that the encoder's hidden representations serve as useful estimations of terminology constraints. It also highlights the robustness of a main-stay data selection approach in the absence of curated IT-related parallel data.

9 Limitations

However, there are some limitations, which open avenues for future work.

The approach to ContexTerm relies on pretrained models to extract the encoders' final hidden states from the source language (or target language). While this makes practical use of existing models, it assumes that such a model exists but this may not be so, especially when there are low resource languages that are under-served.

The current system also does not consider post-processing by lexical constrained decoding. Its interaction with ContexTerm might result in overall translation quality and Terminology Success Rates.

The general translation quality is still not up to par with state-of-the-art neural machine translation engines, and this could be due to severe domain mismatches. As mentioned previously, carefully curated IT-related parallel data is under-resourced. Although CED was used to select domain-specific data from a large pool of domain-agnostic data, this approach may not actually capture actual human-curated IT-related data. This is especially true for German, whose general translation quality falls below 20.

More IT-related parallel data is needed and it would be interesting to see how the contextual approach would fair against NMT systems that have been trained on quality IT-related data.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Lingua custodia's participation at the WMT 2021 machine translation using terminologies shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 799–803, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Amittai Axelrod, Mei Yang, Kevin Duh, and Katrin Kirchhoff. 2008. The university of washington machine translation system for acl wmt 2008. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, page 123–126, USA. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Lynne Bowker. 2021. Machine translation literacy instruction for non-translators: A comparison of five delivery formats. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 25–36, Held Online. INCOMA Ltd.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout, and Raheel Qadar. 2023. Lingua custodia's participation at the WMT 2023 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 897–901, Singapore. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. *Preprint*, arXiv:1903.08788.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook fair's wmt19 news translation task submission. *Preprint*, arXiv:1907.06616.

Tommi Nieminen. 2023. OPUS-CAT terminology systems for the WMT23 terminology shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 912–918, Singapore. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, and Yeonsoo Lee. 2023. VARCO-MT: NCSOFT's WMT'23 terminology shared task submission. In *Proceedings of the Eighth Conference on Machine Translation*, pages 919–925, Singapore. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2024. Evaluation and large-scale training for contextual machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1125–1139, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Adriaan M. J. Schakel and Benjamin J. Wilson. 2015. Measuring word significance using distributed representations of words. *Preprint*, arXiv:1508.02297.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Yu Yuan, Yuze Gao, Yue Zhang, and Serge Sharoff. 2018. Cross-lingual terminology extraction for translation quality estimation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A NMT Training Configurations

Hyperparameters	**en-de**	**en-es**	**en-ru**
Encoder Embedding Size	512	512	512
Decoder Embedding Size	512	512	512
Encoder FFN Embedding Size	2048	2048	2048
Decoder FFN Embedding Size	2048	2048	2048
Encoder Layers	6	2	6
Decoder Layers	6	2	6
Encoder Attention Heads	16	16	16
Decoder Attention Heads	16	16	16
Learning Rate	3e-4	5e-4	3e-4
lr scheduler	Inverse Sqrt	Inverse Sqrt	Inverse Sqrt
Optimizer	Adam	Adam	Adam
Max Tokens	10000	10000	10000
Attention Dropout	0.0	0.0	0.0
Dropout	0.2	0.3	0.3
Criterion	Label-Smoothed CE	Label-Smoothed CE	Label-Smoothed CE
Warmup Steps	8000	6000	8000
Warmup init lr	1e-8	1e-8	1e-8
Clip Norm	1.0	1.0	1.0
Label Smoothing	0.1	0.1	0.1
Max source-target positions	4096	4096	4096

Table 4: Language-pair-specific hyperparameters that were used for NMT training

IRB-MT at WMT25 Terminology Translation Task: Metric-guided Multi-agent Approach

Ivan Grubišić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
ivan.grubisic@irb.hr

Damir Korenčić*
Division of Electronics
Ruđer Bošković Institute
Zagreb, Croatia
damir.korencic@irb.hr

Abstract

Terminology-aware machine translation (MT) is needed in case of specialized domains such as science and law. Large Language Models (LLMs) have raised the level of state-of-the-art performance on the task of MT, but the problem is not completely solved, especially for use-cases requiring precise terminology translations. We participate in the WMT25 Terminology Translation Task with an LLM-based multi-agent system coupled with a custom terminology-aware translation quality metric for the selection of the final translation. We use a number of smaller open-weights LLMs embedded in an agentic "translation revision" workflow, and we do not rely on data- and compute-intensive fine-tuning of models. Our evaluations show that the system achieves very good results in terms of both `MetricX`-24 and a custom `TSR` metric designed to measure the adherence to predefined term mappings.

1 Introduction

When translating texts from specialized technical domains such as medicine, finance, or law, it is important to translate technical terms accurately and consistently (Castilho and Knowles, 2025; Oncevay et al., 2025). To this end, the translation systems can be provided with an existing list of terms and their translations. While the LLMs have emerged as state-of-the-art models for MT (Kocmi et al., 2024), they are rarely evaluated on specialized domains that require strict adherence to terminology. The goal of the WMT25 Terminology Translation Task (Semenov et al., 2025) is to determine how well do the modern MT systems tackle this challenge.

Recently, a number of capable multilingual LLMs that support instruction following were made available to the community (Team et al., 2025; Yang et al., 2025; Martins et al., 2025). In parallel, research in multi-agent workflows showed that embedding individual LLMs in multi-step workflows leads to performance gains. These workflows can be generic, such as self-refine (Madaan et al., 2023), or task-oriented workflows in which agents are assigned natural task-specific roles (Wu et al., 2024; Briakou et al., 2024).

Our goal was to propose a resource-efficient solution based on smaller instruction- and reasoning-capable multilingual LLMs with open weights, embedded in an agentic workflow for performance improvement. We hypothesize that such a workflow could lead to solid performance for a number of language pairs, as the multilinguality of modern LLMs facilitates translation, and their instruction-following capabilities enable the implementation of complex terminology- and revision-related instructions. Such a system does not require datasets and compute for model adaptation and fine-tuning – only relatively modest inference-time compute to run the agents is needed. Participation in the WMT25 Terminology Task enables us to compare such a system with various other approaches.

WMT25 Terminology Task datasets are divided into `Track1` and `Track2`, consisting of texts from the information technology and financial domains, respectively (Semenov et al., 2025). `Track1` datasets contain paragraph-level texts and cover en-de, en-es, and en-ru language pairs. `Track2` datasets contain long document with document-level terminology mappings and cover en-zh and zh-en pairs. Predefined source→translation term mappings are included in the datasets and they come in two flavors: "proper" terminologies covering technical terms, and "random" terminologies with random words. The idea is to measure the influence of the predefined terminology on the performance. For the same reason, an additional "no terminology" setup is included in the task.

Our system, named `MeGuMa`, is an agentic translation system that operates in three phases: 1) translation with individual LLMs; 2) translation revision

*Equal contribution.

Proceedings of the Tenth Conference on Machine Translation, pages 1302–1334
November 8-9, 2025

based on reasoning LLMs; 3) the selection phase using a custom terminology-aware translation quality metric. Revision agents act as senior translators who revise the work of junior translators (individual LLMs). In the revision phase, a reviser LLM examines a number of phase one translations and composes the final translation. Variants of Gemma3, Qwen3, and EuroLLM models are employed as base translators, while Gemma3 and Qwen3 models set up for "thinking" mode are used for revision. The final translation is selected from all the generated translations using a combination of two translation measures: MetricX (Juraska et al., 2024), and a custom TSR metric that measures the adherence to predefined source→translation term mappings.

The final system achieves good results. For Track1 all MetricX scores (range between 0 and 25, lower is better) are below 2.00, while the TSR scores are above 0.85 (indicating that 85% of source terms are correctly translated). For Track2 all MetricX scores are below 2.5 (below 2.0 for the "proper terms" setup), and the TSR scores range between 0.69 and 0.80 except for the random terms setup with scores between 0.40 and 0.50.

Our contributions are: a new agentic system for terminology-aware machine translation, a new TSR metric for approximating the adherence to predefined terminology, and a TTQ metric that aggregates MetricX and TSR in order to combine translation quality and the adherence to predefined terminology. We make the code and the models' output freely available.[1]

2 Datasets

Three different collections of datasets were released for the Terminology MT task: 1) DEV, a set of three datasets for sentence- and paragraph-level terminology translation task; 2) Track1, a set of three datasets for sentence- and paragraph-level terminology translation task; and 3) Track2, a set of ten test datasets for document-level terminology translation task. Each of these two tracks focus on a different domain, with Track1 dealing with the *information technology* domain, while Track2 consists of texts and terminology from the *finance* domain (Semenov et al., 2025).

The datasets in DEV and Track1 are quite similar, with the same language pairs and one unique dataset per pair: English to German, English to Spanish, and English to Russian. Track2 contains ten unique datasets – five datasets for English to Traditional Chinese translation and five datasets for Traditional Chinese to English.

Each text from DEV, Track1, and Track2 datasets is associated with a set of terms and their translations. These term mappings define the terminology-aware translation tasks. For DEV and Track1 a term dictionary is provided at the sentence/paragraph level, and it contains source_term : target_term entries. For Track2 each document is provided with a dictionary of source terms mapped to a list of viable target terms. Each Track1 and Track2 has three variants of term mappings: 1) no terms; 2) proper terms; and 3) random terms. These three variants differ in their inputs. The *no terms* variant uses only input texts, the *proper* terminology variant adds specialized dictionaries for domain-specific terms, and the *random* terminology variant uses dictionaries with randomly selected words from the texts to compare the impact of accurate versus arbitrary terminology on system performance (Semenov et al., 2025).

While the DEV and Track1 datasets contain 500 smaller texts with an average of approximately 10 whitespace-separated tokens (see the Appendix, Table 9, 10, 11 and 12), each Track2 dataset contains between 9 and 13 long texts with approximately 50 paragraphs per text, where each paragraph has approximately 50 tokens on average (see the Appendix, Table 16, 14 and 15).

To reduce the required computational resources and increase the quality of translations and evaluations, we convert Track2 translation datasets from the document-level to the paragraph-level format of Track1. This process requires two steps: text splitting and term splitting. Following a data analysis showing that Track2 texts contain paragraphs separated by two or more newline characters, we first use regular expressions to split texts into paragraphs. However, since the term mappings are provided on the document-level, another step is needed to create paragraph-level term dictionaries containing only the source terms that occur in a paragraph.

To achieve this, for each paragraph we examine all source terms in the document-level term dictionary and check if a source term is present in the paragraph text. However, we need to prevent subterms (substrings of longer terms) to be detected in places where their super-term exists. To this end

[1] https://github.com/igrubi/irb-mt-wmt2025

we employ these steps: 1) lowercase the text and all terms; 2) sort the list of all document terms by decreasing length (longest terms at the start of the list); 3) iterate through the term list and check if a term is present in the paragraph text. If a term is detected, we save it to the paragraph-level dictionary and remove all occurrences of the term from the paragraph text. This way a sub-term will not be included after a super-term. With this procedure we create a paragraph-level term dictionary that contains only a relevant subset of document-level terms.

With this procedure we only reduce the set of dictionary keys to the paragraph-level. The dictionary values remain the same as in the original `Track2` datasets, i.e., each source term is still mapped to a list of viable target terms. This is the key difference between the generated `Track2` paragraph-level term dictionaries and the `Track1` dictionaries, and we tackle these two cases differently with translation and revision agents.

Finally, in order to recreate the whole translated document later, we give each paragraph a unique identifier that contains information on the exact location of the paragraph within the document.

After the final submission we discovered that there was a bug in the function that we used for splitting Chinese terms. Namely, when string-matching a term in a text, we required the term to be surrounded by word-separating whitespace. While this is a proper approach for European languages, it is not for Traditional Chinese because often there is no such separation between words. As a result, we lost a good number of the paragraph terms for the datasets where the source language is Chinese. For details, see the Appendix, Subsection A.3, difference between Table 16 and 17.

3 The System

Our translation system produces an output in three phases: 1) individual translator LLMs generate initial translations, 2) reviser LLMs generate improved translations from the initial ones, and 3) all of the candidate translations are pooled, and the best one is selected based on a custom quality metric.

In the development phase we benchmarked a number of translator-LLM candidates on a development set consisting of a subset of pairs from both the `DEV`, `Track1` and `Track2` datasets. Most of the evaluated LLMs demonstrated good results and were included into the final system.

We start the detailed description of the complete system with metrics, used both for the benchmarking of LLM and for the selection of final system outputs.

3.1 Metrics

General translation quality. To evaluate the general (not terminology-aware) translation quality, we use `MetricX` metric (Juraska et al., 2024) (the "metricx-24-hybrid-xl-v2p6" variant). `MetricX` was chosen since its scores of translation quality are highly correlated with human judgments (Juraska et al., 2024), and it does not require a reference translation in order to assess translation quality. This makes `MetricX` applicable for assessing translation performance on `Track1` and `Track2` test datasets.

Terminology translation success rate. When performing terminology-aware translation, it is important that the terms from the source language are translated to the target language according to the provided dictionary with correct term mappings. To measure this we designed a custom metric that does not require a reference translation and is therefore usable both for benchmarking on the test set and for filtering multiple generated translations.

The metric, dubbed the Terminology Success Rate (`TSR`) metric, directly compares source-language terms present in the source text with the target-language terms present in the translation. The input to the metric consists of the source and target texts, and of the dictionary of term translations. The metric relies on lemmatizers and sentence segmenters for both source and target languages, and a sentence-level alignment module.

In the first step, source and target texts are segmented into sentences. In the second step two sets of sentences are aligned using the SentAlign method that relies on a multilingual sentence embedding module and an alignment optimization algorithm (Steingrimsson et al., 2023). The output of SentAlign is a list of pairs of matching sentence blocks – one or more source sentences can be aligned to one or more target sentences, and a sentence can be without a match. Given a pair of aligned (sub)texts, and a (source_term, target_term) pair from the dictionary such that the source term appears in the source text, a pair-level score is calculated as the percentage of matched target terms in relation to the matched source terms. This score,

capped to 100%, approximates the coverage of the source terms occurring in source text by the target terms occurring in the translation. The final score is calculated by averaging all per-term scores for each pair of aligned texts, and then averaging over all pairs in the alignment.

Matching of terms to their occurrences in texts is preceded by lowercasing and lemmatization in order to account for different surface forms that a term can assume. To avoid over-counting of terms the matching process takes into account the term overlap – longer terms are matched first and each subsequent term is matched only to a positions where it does not overlap with any previously matched term.

Terminology translation quality. Both previously described translation metrics are equally important to assess the final terminology-aware translation quality. For this reason we combine these metric into a new metric that we name Terminology Translation Quality (TTQ).

To achieve this, we convert MetricX from the 25-0 scale (where a lower score means better translation) to the 0-1 scale (where a higher score means better translation), using the following equation:

$$\tilde{\text{MetricX}} = (25 - \text{MetricX})/25$$

Then the final TTQ metric is calculated by averaging the arithmetic, geometric, and harmonic means of converted MetricX and TSR.

While the arithmetic mean is influenced by both scores equally ignoring the relation between their values, the harmonic and geometric mean progressively penalize cases with larger differences giving more weight to smaller values. Since we want the TTQ metric to capture both of these characteristics, we define it as the average of all three means. In this way, we enforce the discriminatory strength of the arithmetic mean in those cases where one score is extremely small. The TTQ metric produces a stable and fair score capable of discriminating between similar solutions.

3.2 LLM Benchmarking and Selection

Translation is performed by LLM-based agents in two phases - *translation* and *revision*. We use pretrained, relatively small (8-27 billion parameters) and open-weights LLMs. No further adaptation or fine-tuning is performed, and no additional data is used.

As candidates for the translation we considered the following LLMs: Gemma3 (12B and 27B) (Team et al., 2025), Qwen3 (8B and 14B – both thinking and non-thinking variants) (Yang et al., 2025), and EuroLLM (9B) (Martins et al., 2025). These are recent LLMs, built using state-of-the-art approaches, that have multilingual and instruction-following capabilities.

As a first step, we performed a test of basic translation capabilities on a subsample of DEV and Track2 datasets. We took 10 longest texts from each DEV dataset, and 10 longest paragraphs for each Track2 dataset, and evaluated the LLMs with both MetricX and TSR. The models and their variants tested were: gemma3_27b, gemma3_12b, qwen3_14b-think, qwen3_8b-think, qwen3_14b, qwen3_8b, and eurollm_9b.

The results, displayed in Tables 1 and 2, show that all models have competitive performance for almost all language pairs. Some of the models appear to be biased towards MetricX, like eurollm_9b, some others are more inclined to TSR, like qwen3_8b-think, while some models are balanced between these two metrics, like gemma3_27b. Only eurollm_9b shows slightly poorer translation performance from English to Chinese and vice versa in both metrics (Table 2). For more details, see the Appendix, Section D.

Based on these results we decided to use all the model variants except eurollm_9b as basic translators for both Track1 and Track2. We only use eurollm_9b as one of the translators for the European languages of Track1.

The idea of a revision phase comes from the translation-revision system, which is a successful practice performed by professional translators for years (Arthern, 1978). Although basic translation can be carried out by less experienced personnel, the revision requires an experienced translator to produce final high-quality translations of the texts. This is the reason why we speculated that the revision agent should be a larger and more capable LLM. This intuition was confirmed by the pilot experiment that we conducted on the subsample of the DEV and Track2 datasets, which showed that reviser agents based on smaller models, such as revis-qwen3_8b-think and revis-eurollm-think, produce a large number of errors.

Therefore, we decided to use as revis-

agents	ende			enes			enru		
	MetricX	TSR	TTQ	MetricX	TSR	TTQ	MetricX	TSR	TTQ
trans-eurollm	**1.17**	0.44	0.65	**3.39**	0.78	0.82	3.10	0.44	0.62
trans-qwen3_8b	1.36	0.56	0.73	3.67	0.88	0.86	4.77	0.68	0.74
trans-qwen3_8b-think	1.80	0.62	0.76	3.53	0.93	0.89	4.50	0.77	0.80
trans-qwen3_14b	1.37	0.56	0.73	3.53	0.90	0.88	3.03	0.69	0.78
trans-qwen3_14b-think	1.23	0.55	0.72	3.86	**0.95**	**0.90**	3.82	**0.85**	**0.85**
trans-gemma3_12b	1.61	0.64	0.77	3.52	**0.95**	**0.90**	3.00	0.74	0.81
trans-gemma3_27b	1.36	**0.75**	**0.84**	3.49	0.93	**0.90**	**2.66**	0.80	**0.85**

Table 1: Combined mean scores from English to German (ende), Spanish (enes), and Russian (enru), showing MetricX, TSR and TTQ scores evaluated on subset of DEV datasets.

agents	enzh			zhen		
	MetricX	TSR	TTQ	MetricX	TSR	TTQ
trans-eurollm	4.06	0.30	0.50	3.50	0.50	0.65
trans-qwen3_8b	3.82	0.40	0.58	3.22	0.51	0.67
trans-qwen3_8b-think	3.90	0.41	0.59	3.26	**0.53**	**0.68**
trans-qwen3_14b	**3.51**	0.41	0.59	3.34	0.52	0.67
trans-qwen3_14b-think	**3.51**	0.41	0.59	**3.20**	0.52	0.67
trans-gemma3_12b	4.40	**0.45**	**0.61**	3.31	0.51	0.66
trans-gemma3_27b	3.93	**0.45**	**0.61**	3.43	0.52	0.67

Table 2: Combined mean scores for Track2 subset translating English to Traditional Chinese (enzh) and Traditional Chinese to English (zhen), showing MetricX, TSR and TTQ scores evaluated on subset of Track2 datasets.

ers only agents based on the largest variants of Gemma3 and Qwen3 models that showed solid performance in all languages: revis-gemma3_27b-think, revis-gemma3_12b-think, and revis-qwen3_14b-think. Unlike the standard translation-revision system where the reviser receives only one translation, our revision agents receive as input translations from all of the basic translation agents, and produce their final translations from the entire input. To increase the reasoning power, and thus the quality of revised texts, all of the revision models were prompt-induced to first think before producing the final solution. Similar to *chain-of-thought* prompting (Wei et al., 2022).

3.3 Context Engineering

Our system is based on prompt-guided LLM agents. Therefore, the output of these agents is highly dependent on the context of the task, i.e. the way the context is compiled and formatted for the LLM.

As there are three different cases of term data in the test datasets, we created an appropriate prompt for each case: 1) Case 1, when no term data is provided; 2) Case 2, when a single translated term is given for each source term (Track1); and 3) Case 3, when multiple viable translated terms are provided for each source term (Track2). The final prompt versions are listed in the Appendices: for *translation prompts* look at B and for *revision prompts* at C.

Each prompt consists of two parts: the system prompt and the user prompt. For translation-only agents, the system prompt defines a role of the agent, explains the task, describes the key requirements of the task, and provides the term dictionary if the term data is available. The user prompt provides the text that is to be translated.

For the reviser agents the system prompt is quite similar to the translation-only one. However, it is expanded with the original source text and with all translations produced by the translation agents. The user prompt only gives a short summary of the task. An additional difference from the translation-only prompt is the "thinking command". This command is used to induce the agent to think about the given translations and about the potential enhancements prior to giving the final revised translation.

3.4 The Final Translation System

Our system is based on multi-agent approach with three steps: 1) translation; 2) revision; 3) selection. The translation step is performed by a number of agents based on open-weight LLMs that each translate an input text from a source to a target language. The revision step is performed by three LLM agents. Each reviser agent receives all the translations from the translation step and produces a revised translation. In the selection step we use the TTQ metric to evaluate all of the produced translations (from both the translation and the revision step), and select the best one as the final system output. Additionally, to enable a unified approach for both tracks, we perform a data preparation step in which we break Track2 document-level texts into paragraph-level texts corresponding to the texts of Track1. During this process document-level terminology dictionary is projected to each of the paragraphs. The details of the process are explained in Section 2.

4 Results

The final system produces high-quality translations with low MetricX scores (Tables 3 and 5) and high TSR scores (Tables 4 and 6). TSR scores are significantly higher for Track1 than for Track2– this could be caused by both the language differences (European languages vs. Chinese) and by differences in the complexity of the terminology dictionaries (single-choice vs. multi-choice terms).

pair	noterm	proper	random	mean
ende	0.36	0.89	0.76	0.67
enes	1.24	1.73	1.63	1.53
enru	0.77	1.38	1.23	1.13

Table 3: MetricX scores of final solution for Track1.

pair	proper	random	mean
ende	0.86	0.87	0.87
enes	0.88	0.88	0.88
enru	0.94	0.93	0.94

Table 4: TSR scores of final solution for Track1.

When compared to the translation produced by all of the LLM agents used (both translators and revisers) the final system's translations have the best MetricX and TSR scores for almost all Track1 and Track2 datasets. The only exceptions are: 1) translations from English to German for the "proper term" case where a standalone trans-eurollm has the best MetricX score (Track1, see the Appendix, Table 23); and 2) translations from English to Traditional Chinese for the "random term" case where trans-qwen3_14b has the best MetricX scores (Track2, see the Appendix, Table 30).

The final system uses a metric-guided approach based on the custom TTQ metric used to select the best output from a pool of translations. The selection step is an important part of the proposed translation pipeline since it significantly increases translation quality scores, raising them above the scores of the best individual translators and the scores of the best reviser agents (see the Appendix, Section E). The reason why the selector chooses the best translation from the pool of translations of all agents and not only the reviser agents is shown in Tables 7 and 8. These tables show the frequency with which an agent's output was chosen for the final solution. Although the reviser agents have a higher chance of being chosen by the selector (17.13% vs. 6.89% in Track1, and 14.3% vs. 9.42% in Track2), in total about half of all samples are selected from individual translator agents (48.2% in Track1 and 56.5% in Track2).

In addition, there is an interesting trend in the results. MetricX scores are increasing (i.e. translation quality drops) for random and proper term cases compared to the no-term case (Tables 3 and 5; for more information, see the Appendix, Subsections E.1.1, E.2.1, E.2.2 and E.2.3). We hypothesize that forcing a predefined translation of a set of terms reduces the general translation quality because translators need to balance between two different objectives: 1) general translation and 2) terminology constraint. However, this hypothesis requires more comprehensive research to be proven.

Upon submission, our system was evaluated[2] and compared to other participating systems (20 systems in Track1 and 4 systems in Track2) (Semenov et al., 2025). ChrF2++ was used to measure translation quality and a custom terminology success rate metric (which we label Term-Acc) was used to measure adherence to the predefined terminology (Semenov et al., 2025).

[2] https://github.com/wmt-conference/wmt25-terminology/

agents	2015	2017	2019	2021	2023	mean
enzh.noterm	1.49	1.40	1.37	1.41	1.46	1.43
enzh.random	2.39	2.28	2.26	2.23	2.22	2.28
enzh.proper	1.88	1.70	1.73	1.77	1.84	1.78
zhen.noterm	1.22	1.17	1.12	1.14	1.14	1.16
zhen.random	1.54	1.58	1.53	1.46	1.52	1.53
zhen.proper	1.59	1.57	1.47	1.58	1.53	1.55

Table 5: `MetricX` scores of final solution for `Track2`.

agents	2015	2017	2019	2021	2023	mean
enzh.random	0.75	0.70	0.72	0.73	0.75	0.73
enzh.proper	0.78	0.77	0.78	0.81	0.79	0.79
zhen.random	0.39	0.48	0.48	0.48	0.46	0.46
zhen.proper	0.69	0.74	0.75	0.80	0.75	0.75

Table 6: `TSR` scores of final solution for `Track2`.

agents	ende	enes	enru	mean
`trans-eurollm`	9.7%	12.4%	11.0%	11.0%
`trans-qwen3_8b`	5.4%	5.1%	4.9%	5.1%
`trans-qwen3_8b-think`	5.2%	6.3%	6.2%	5.9%
`trans-qwen3_14b`	6.2%	7.0%	6.0%	6.4%
`trans-qwen3_14b-think`	4.8%	4.0%	4.0%	4.3%
`trans-gemma3_12b`	8.6%	6.5%	9.8%	8.3%
`trans-gemma3_27b`	6.8%	7.3%	7.4%	7.2%
`revis-qwen3_14b-think`	11.5%	10.0%	10.6%	10.7%
`revis-gemma3_12b-think`	**24.8%**	**25.3%**	**24.8%**	**24.9%**
`revis-gemma3_27b-think`	16.6%	15.8%	15.1%	15.8%

Table 7: Aggregated total frequencies of agents selected for the final solution across English to German, Spanish, and Russian translations (`Track1`).

agents	enzh-n	enzh-p	enzh-r	zhen-n	zhen-p	zhen-r	mean
`trans-qwen3_8b`	7.4%	6.0%	8.6%	8.6%	9.7%	8.0%	8.1%
`trans-qwen3_8b-think`	6.1%	6.4%	7.3%	7.1%	6.0%	7.3%	6.7%
`trans-qwen3_14b`	6.8%	6.8%	7.0%	9.4%	11.3%	10.7%	8.7%
`trans-qwen3_14b-think`	8.7%	8.2%	7.1%	9.5%	8.5%	8.0%	8.3%
`trans-gemma3_12b`	11.6%	14.4%	15.8%	15.2%	9.8%	12.6%	13.2%
`trans-gemma3_27b`	10.6%	11.7%	15.4%	9.7%	10.5%	11.4%	11.5%
`revis-qwen3_14b-think`	12.6%	11.7%	8.2%	**15.8%**	**17.4%**	**16.0%**	13.6%
`revis-gemma3_12b-think`	**17.3%**	**20.0%**	**16.1%**	13.5%	14.8%	12.6%	**15.7%**
`revis-gemma3_27b-think`	18.4%	14.3%	14.1%	10.7%	11.3%	13.0%	13.6%

Table 8: Aggregated total frequencies of agents selected for final solutions across different translation directions and term conditions (`Track2`).

In Track1, our system has an average ChrF2++ of 67.2 (6th place) and a high average Term-Acc of 97.4 (4th place). In terms of Pareto optimality between ChrF2++ and Term-Acc, our system is near-optimal, with only two systems having Pareto dominance over it: o3-term-guide and duterm.

In Track2, our system has an average ChrF2++ of 54.3 (3rd place) and a competitive average Term-Acc of 79.5 (2nd place), with only one system, CommandA_WMT, having Pareto dominance over it.

In our final submission there is an error for the proper and random cases of Track2 translations from Traditional Chinese to English. The error originates in the data preparation process (more information can be found at the end of Subsection 2). We estimated the drop in scores as the result of this error (see the Appendix, Subsections E.2.4 and E.2.8). The expected drop is around 0.30 MetricX scores and around 0.03 and 0.11 TSR scores for random and proper term cases.

Expectedly, this error effects the official results of our system: there is a large gap between the Term-Acc score for en-zh (96.6) and the Term-Acc score for zh-en (62.4) (Semenov et al., 2025). To examine the impact of the error, we evaluated[3] the debugged version of the system and obtained the zh-en Term-Acc of 96.6 (see E.4 for more details). Our repository[4] contains details of the error, reproducibility instructions, and the outputs of both error-containing and error-free versions of the system.

Conclusion and Future Work

Our system uses a multi-agent approach with three steps: 1) translation; 2) revision; 3) selection. The idea of a revision comes from the translation-revision system, which is a successful practice carried out by professional translators for years (Arthern, 1978). Here, we expand this practice that uses one human translator and one human reviser to multiple LLM translators and multiple LLM revisers. In addition, we use metric-guided selection as the final step of our system workflow. Here, we evaluate each translation with the custom TTQ metric and select the best one as the final output. Finally, the final solution yields the best performance compared to all translations and revisions for all language pairs and terminology constraint cases.

Internal evaluations show that the proposed system has very good translation scores in terms of both metrics: MetricX (which measures the general translation quality) and TSR (which measures the terminology success rate).

The contributions of this work are a novel agentic approach for terminology-aware machine translation, a novel TSR metric for measuring the adherence to the predefined term translations, and a novel TTQ metric that aggregates MetricX and TSR in a score that combines translation quality and correctness of terminology translation.

Future work will focus on expanding the evaluation to more language pairs in order to test the robustness of the system. Additionally, evaluation on translation datasets that contain reference translations would expectedly provide a more precise assessment of the system's performance.

Qualitative analysis of system outputs, based on evaluations by humans or by the top LLM systems, could lead to valuable insights and improvements. Measuring how much the system improves the productivity of translators in a real-world setting would also be valuable. To enable a more fine-grained analysis of the we make available the outputs of all the constituent LLMs.[4]

The TSR metric should be evaluated against human quality scores on a diverse set of languages pairs in order to verify its quality and robustness. The assessment of the quality and reliability of the TTQ metric also requires validation against human annotations.

Finally, we plan to improve our system with a more granular agentic workflow that incorporates additional specialized roles like pre-editor and post-editor. The key challenge of such improvements is boosting performance while reducing the execution time, influenced by both the number of workflow steps and by the size of the LLMs used.

Acknowledgements

This paper was supported by the European Union's NextGenerationEU program. We would like to thank Tomislav Šmuc, Ph.D., and Prof. Sonja Grgić, Ph.D., for support and valuable discussions. We acknowledge EuroHPC Joint Undertaking for awarding us access to MareNostrum5 hosted by BSC, Spain, under the project ID EHPC-DEV-2025D05-087.

[3] https://github.com/wmt-conference/wmt25-terminology/

[4] https://github.com/igrubi/irb-mt-wmt2025

References

Peter J. Arthern. 1978. Machine translation and computerised terminology systems - a translator's viewpoint. In *Translating and the Computer*, London, UK. Aslib Proceedings.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Sheila Castilho and Rebecca Knowles. 2025. A survey of context in neural machine translation and its evaluation. *Natural Language Processing*, 31(4):986–1016.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2758–2775, Albuquerque, New Mexico. Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy

Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

A Data analysis

A.1 DEV

DATASET	TEXTS
ende	500
enes	500
enru	500

Table 9: Number of texts in DEV. datasets

DATASET	MIN	MEAN	MAX	TOTAL
ende	2	9.98	49	4991
enes	3	10.79	46	5397
enru	2	9.43	50	4716
ALL	2	10.07	50	15104

Table 10: Number of tokens in texts in DEV datasets.

A.2 Track1

DATASET	TEXTS
ende	500
enes	500
enru	500

Table 11: Number of texts in Track1 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
ende.noterm	2	10.40	50	5199
enes.noterm	3	10.60	41	5298
enru.noterm	3	9.02	44	4509
ALL	2	10.00	50	15006

Table 12: Number of tokens in texts in Track1 datasets.

A.3 Track2

DATASET	DOCUMENTS
2015.enzh	9
2016.zhen	10
2017.enzh	10
2018.zhen	10
2019.enzh	11
2020.zhen	11
2021.enzh	12
2022.zhen	12
2023.enzh	13
2024.zhen	13

Table 13: Number of documents in Track2 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2015.enzh	18	38.11	68	343
2016.zhen	17	37.00	63	370
2017.enzh	19	43.10	119	431
2018.zhen	19	41.60	91	416
2019.enzh	24	40.27	71	443
2020.zhen	21	46.64	121	513
2021.enzh	28	47.08	119	565
2022.zhen	23	43.25	111	519
2023.enzh	14	41.62	107	541
2024.zhen	15	44.62	91	580
ALL	14	42.53	121	4721

Table 14: Number of paragraphs per document in Track2 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2015.enzh	1	43.00	232	14748
2016.zhen	2	76.21	559	28199
2017.enzh	1	43.26	366	18644
2018.zhen	2	72.33	521	30089
2019.enzh	1	44.37	301	19654
2020.zhen	2	67.22	559	34485
2021.enzh	1	39.51	340	22323
2022.zhen	2	69.03	559	35827
2023.enzh	1	42.42	340	22948
2024.zhen	2	64.67	374	37510
ALL	1	56.01	559	264427

Table 15: Number of tokens per paragraph in Track2 datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2015.enzh.proper	0	3.86	33	1324
2015.enzh.random	0	8.56	52	2936
2016.zhen.proper	0	3.20	22	1182
2016.zhen.random	1	21.94	114	8118
2017.enzh.proper	0	3.27	18	1409
2017.enzh.random	0	8.65	59	3727
2018.zhen.proper	0	3.77	21	1569
2018.zhen.random	1	22.6	122	9411
2019.enzh.proper	0	3.01	19	1333
2019.enzh.random	0	9.71	62	4302
2020.zhen.proper	0	3.22	20	1650
2020.zhen.random	0	20.02	92	10269
2021.enzh.proper	0	3.63	16	2050
2021.enzh.random	0	8.14	52	4601
2022.zhen.proper	0	3.18	17	1648
2022.zhen.random	0	20.15	83	10458
2023.enzh.proper	0	3.64	18	1970
2023.enzh.random	0	8.02	49	4340
2024.zhen.proper	0	3.23	16	1871
2024.zhen.random	1	19.91	99	11552
TOTAL	0	9.09	122	85720
ALL	0	3.26	62	30787

Table 16: Number of terms per paragraph in `Track2` datasets.

DATASET	MIN	MEAN	MAX	TOTAL
2016.zhen.proper	0	0.36	6	134
2016.zhen.random	0	1.01	10	372
2018.zhen.proper	0	0.51	7	214
2018.zhen.random	0	0.68	11	283
2020.zhen.proper	0	0.43	6	220
2020.zhen.random	0	0.64	8	330
2022.zhen.proper	0	0.58	7	300
2022.zhen.random	0	0.54	6	281
2024.zhen.proper	0	0.52	11	301
2024.zhen.random	0	0.62	8	360
ALL	0	3.26	62	30787

Table 17: Number of terms per paragraph in `Track2` datasets affected by data preparation bug.

B Translation prompts

B.1 Case 1 - no terms prompt

B.1.1 System prompt:

You are a professional translator specializing in **{source_language}** to **{target_language}**.

Your task is to translate the provided **{source_language}** text into fluent and natural **{target_language}**.

Key requirements:

- Accurately convey the meaning and nuances of the original text, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- Provide only the full **{target_language}** translation as output. Do not include any explanations, comments, or additional text.

B.1.2 User prompt:

Translate the following **{source_language}** text into **{target_language}**:
{text}

B.2 Case 2 - single-choice terms prompt (Track1)

B.2.1 System prompt:

You are a professional translator specializing in **{source_language}** to **{target_language}**.

Your task is to translate the provided **{source_language}** text into fluent and natural **{target_language}**.

Key requirements:

- Accurately convey the meaning and nuances of the original text, respecting the grammar, vocabulary, and cultural norms of **{target_language}**.
- Whenever a **{source_language}** term matches an entry in the dictionary below, replace it with the exact **{target_language}** translation from the dictionary.
- Translate all other text normally, without altering any words not found in the dictionary.
- Provide only the full translation in **{target_language}** as output. Do not include any explanations, comments, or additional text.

Dictionary:
{terms}

B.2.2 User prompt:

Translate the following **{source_language}** text into **{target_language}**:
{text}

B.3 Case 3 - multi-choice terms prompt (Track2)

B.3.1 System prompt:

You are a professional translator specializing in **{source_language}** to **{target_language}**.

Your task is to translate the provided **{source_language}** text into fluent and natural **{target_language}**.

Key requirements:

- Accurately convey the meaning and nuances of the original text, respecting the grammar, vocabulary, and cultural norms of **{target_language}**.
- For any term in the **{source_language}** text that matches a key in the provided dictionary, use exactly one translation from that term's list (choose the best fitting translation in context).
- Translate all other text normally, without altering any words not found in the dictionary.
- Provide only the full translation in **{target_language}** as output. Do not include any explanations, comments, or additional text.

Dictionary:
{terms}

B.3.2 User prompt:

Translate the following **{source_language}** text into **{target_language}**:
{text}

C Revision prompts

C.1 Case 1 - no terms prompt

C.1.1 System prompt:

You are a professional senior translator specializing in **{source_language}** to **{target_language}**.

You will be given an original text in **{source_language}** followed by several translations into **{target_language}** produced by junior translators.

Your first task: Review the provided translations with these requirements:

- Critically evaluate each translation, noting strengths and weaknesses.
- Focus your observations on translation quality, fluency, grammar, vocabulary, and cultural appropriateness.
- After your review, reason about potential improvements and how to produce the best possible translation.
- Keep your review and reasoning succinct (under 1000 words).
- Enclose your review and reasoning within the **<think>** and **</think>** tags.

Your second task: Translate the original **{source_language}** text into fluent, natural **{target_language}**, following these guidelines:

- Complete this task only after the first task.
- Produce the best possible translation based on your previous reasoning.
- Accurately convey the meaning and nuance of the original, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- Provide only the final translation as output, without explanations or comments.

Original **{source_language}** text:
{text}

Translations by junior translators:

1. Translation by the first junior translator:
   {translations[0]}

2. Translation by the second junior translator:
   {translations[1]}

3. Translation by the third junior translator:
   {translations[2]}

4. Translation by the fourth junior translator:
   {translations[3]}

5. Translation by the fifth junior translator:
   {translations[4]}

6. Translation by the sixth junior translator:
   {translations[5]}

7. Translation by the seventh junior translator:
   {translations[6]}

C.1.2 User prompt:

First, review these translations and reason about producing the best possible translation, enclosing your review in **<think>** and **</think>**.
Then, provide your improved translation of the original **{source_language}** text into **{target_language}**.

C.2 Case 2 - single-choice terms prompt (Track1)

C.2.1 System prompt:

You are a professional senior translator specializing in **{source_language}** to **{target_language}**.

You will be given an original text in **{source_language}** along with a dictionary of terms that must be translated exactly, followed by several translations into **{target_language}** produced by junior translators.

Your first task: Review the provided translations with these requirements:

- Critically evaluate each translation, noting strengths and weaknesses.

- Focus your observations on translation quality, fluency, grammar, vocabulary, and cultural appropriateness.
- Verify that all **{source_language}** terms matching keys in the dictionary below are correctly translated using the **exact {target_language}** equivalents provided.
- After your review, reason about potential improvements and how to produce the best possible translation.
- Keep your review and reasoning succinct (under 1000 words).
- Enclose your review and reasoning within the **<think>** and **</think>** tags.

Your second task: Translate the original **{source_language}** text into fluent, natural **{target_language}**, following these guidelines:

- Complete this task only after the first task.
- Produce the best possible translation based on your previous reasoning.
- Accurately convey the meaning and nuance of the original **{source_language}** text, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- Replace **every** term in the **{source_language}** text found as a key in the dictionary below with its **exact {target_language}** translation from the dictionary.
- Translate all other text normally, without altering words not found in the dictionary.
- Provide only the final translation as output, without explanations or comments.

Original **{source_language}** text:
{text}

Dictionary:
{terms}

Translations by junior translators:

1. Translation by the first junior translator:
   {translations[0]}
2. Translation by the second junior translator:
   {translations[1]}
3. Translation by the third junior translator:
   {translations[2]}
4. Translation by the fourth junior translator:
   {translations[3]}
5. Translation by the fifth junior translator:
   {translations[4]}

6. Translation by the sixth junior translator:
   {translations[5]}

7. Translation by the seventh junior translator:
   {translations[6]}

C.2.2 User prompt:

First, review these translations and reason about producing the best possible translation, enclosing your review in **<think>** and **</think>**.
Then, provide your improved translation of the original **{source_language}** text into **{target_language}**.

C.3 Case 3 - multi-choice terms prompt (Track2)

C.3.1 System prompt:

You are a professional senior translator specializing in **{source_language}** to **{target_language}**.

You will be given an original text in **{source_language}** along with a dictionary of terms that must be translated exactly, followed by several translations into **{target_language}** produced by junior translators.

Your first task: Review the provided translations with these requirements:

- Critically evaluate each translation, noting strengths and weaknesses.
- Focus your observations on translation quality, fluency, grammar, vocabulary, and cultural appropriateness.
- Verify that all **{source_language}** terms matching keys in the dictionary below are correctly translated using one of the **{target_language}** alternatives listed for that term.
- After your review, reason about potential improvements and how to produce the best possible translation.
- Keep your review and reasoning succinct (under 1000 words).
- Enclose your review and reasoning within the **<think>** and **</think>** tags.

Your second task: Translate the original **{source_language}** text into fluent, natural **{target_language}**, following these guidelines:

- Complete this task only after the first task.
- Produce the best possible translation based on your previous reasoning.
- Accurately convey the meaning and nuance of the original **{source_language}** text, respecting **{target_language}** grammar, vocabulary, and cultural norms.
- For each term in the **{source_language}** text found as a key in the dictionary below, replace it with exactly one **{target_language}** translation selected from that term's list (choose the best fitting translation in context).
- Translate all other text normally, without altering words not found in the dictionary.

- Provide only the final translation as output, without explanations or comments.

Original **{source_language}** text:
{text}

Dictionary:
{terms}

Translations by junior translators:

1. Translation by the first junior translator:
   {translations[0]}

2. Translation by the second junior translator:
   {translations[1]}

3. Translation by the third junior translator:
   {translations[2]}

4. Translation by the fourth junior translator:
   {translations[3]}

5. Translation by the fifth junior translator:
   {translations[4]}

6. Translation by the sixth junior translator:
   {translations[5]}

7. Translation by the seventh junior translator:
   {translations[6]}

C.3.2 User prompt:

First, review these translations and reason about producing the best possible translation, enclosing your review in **<think>** and **</think>**.
Then, provide your improved translation of the original **{source_language}** text into **{target_language}**.

D Model selection

D.1 Track1

The model selection for Track1 is done based on the scores achieved on the subset of DEV datasets. The subset contains 10 longest texts from each dataset.

agents	bleu	MetricX	TSR	TTQ
eurollm	0.41	**1.17**	0.44	0.65
qwen3_8b	0.29	1.36	0.56	0.73
qwen3_8b-think	0.30	1.80	0.62	0.76
qwen3_14b	0.39	1.37	0.56	0.73
qwen3_14b-think	0.34	1.23	0.55	0.72
gemma3_12b	0.39	1.61	0.64	0.77
gemma3_27b	**0.43**	1.36	**0.75**	**0.84**

Table 18: Mean scores for DEV subset, translations from English to German.

agents	bleu	MetricX	TSR	TTQ
eurollm	0.45	**3.39**	0.78	0.82
qwen3_8b	0.44	3.67	0.88	0.86
qwen3_8b-think	**0.47**	3.53	0.93	0.89
qwen3_14b	0.45	3.53	0.90	0.88
qwen3_14b-think	0.46	3.86	**0.95**	**0.90**
gemma3_12b	0.44	3.52	**0.95**	**0.90**
gemma3_27b	0.45	3.49	0.93	**0.90**

Table 19: Mean scores for DEV subset, translations from English to Spanish.

agents	bleu	MetricX	TSR	TTQ
eurollm	0.23	3.10	0.44	0.62
qwen3_8b	0.21	4.77	0.68	0.74
qwen3_8b-think	0.26	4.50	0.77	0.80
qwen3_14b	**0.28**	3.03	0.69	0.78
qwen3_14b-think	0.25	3.82	**0.85**	**0.85**
gemma3_12b	0.27	3.00	0.74	0.81
gemma3_27b	0.27	**2.66**	0.80	**0.85**

Table 20: Mean scores for Track2 subset, translations from English to Russian.

D.2 Track2

The model selection for Track2 is done based on the scores achieved on the subset of Track2 datasets. The subset contains 10 longest paragraphs from each dataset.

agents	MetricX	TSR	TTQ
eurollm	4.06	0.30	0.50
qwen3_8b	3.82	0.40	0.58
qwen3_8b-think	3.90	0.41	0.59
qwen3_14b	**3.51**	0.41	0.59
qwen3_14b-think	**3.51**	0.41	0.59
gemma3_12b	4.40	**0.45**	**0.61**
gemma3_27b	3.93	**0.45**	**0.61**

Table 21: Mean scores for Track2 subset, translations from English to Traditional Chinese.

agents	MetricX	TSR	TTQ
eurollm	3.50	0.50	0.65
qwen3_8b	3.22	0.51	0.67
qwen3_8b-think	3.26	**0.53**	**0.68**
qwen3_14b	3.34	0.52	0.67
qwen3_14b-think	**3.20**	0.52	0.67
gemma3_12b	3.31	0.51	0.66
gemma3_27b	3.43	0.52	0.67

Table 22: Mean scores for Track2 subset, translations from Traditional Chinese to English.

E Scores

E.1 Track1

E.1.1 MetricX scores

agents	noterm	proper	random	mean
trans-eurollm	0.70	**0.75**	0.77	0.74
trans-qwen3_8b	0.93	1.34	1.20	1.16
trans-qwen3_8b-think	0.92	1.34	1.27	1.18
trans-qwen3_14b	0.87	1.19	1.04	1.03
trans-qwen3_14b-think	0.86	1.31	1.15	1.11
trans-gemma3_12b	0.70	1.33	1.22	1.08
trans-gemma3_27b	0.73	1.32	1.11	1.05
revis-qwen3_14b-think	1.38	1.89	1.69	1.65
revis-gemma3_12b-think	0.71	1.32	1.12	1.05
revis-gemma3_27b-think	0.69	1.22	1.06	0.99
final	**0.36**	0.89	**0.76**	**0.67**

Table 23: MetricX scores for Track1, translations from English to German.

agents	noterm	proper	random	mean
trans-eurollm	1.83	1.90	1.91	1.88
trans-qwen3_8b	2.04	2.35	2.16	2.18
trans-qwen3_8b-think	1.85	2.31	2.32	2.16
trans-qwen3_14b	1.85	2.11	2.06	2.01
trans-qwen3_14b-think	1.84	2.27	2.25	2.12
trans-gemma3_12b	1.77	2.36	2.35	2.16
trans-gemma3_27b	1.82	2.33	2.20	2.12
revis-qwen3_14b-think	2.53	2.99	2.75	2.76
revis-gemma3_12b-think	1.97	2.30	2.24	2.17
revis-gemma3_27b-think	1.83	2.24	2.13	2.07
final	**1.24**	**1.73**	**1.63**	**1.53**

Table 24: MetricX scores for Track1, translations from English to Spanish.

agents	noterm	proper	random	mean
trans-eurollm	1.66	1.84	1.67	1.72
trans-qwen3_8b	1.77	2.01	2.13	1.97
trans-qwen3_8b-think	1.70	2.28	2.21	2.06
trans-qwen3_14b	1.61	1.96	1.86	1.81
trans-qwen3_14b-think	1.60	2.15	2.09	1.95
trans-gemma3_12b	1.42	2.21	2.11	1.91
trans-gemma3_27b	1.49	2.17	2.20	1.95
revis-qwen3_14b-think	2.25	2.74	2.63	2.54
revis-gemma3_12b-think	1.67	2.18	2.19	2.01
revis-gemma3_27b-think	1.53	2.17	1.97	1.89
final	**0.77**	**1.38**	**1.23**	**1.13**

Table 25: MetricX scores for Track1, translations from English to Russian.

E.1.2 TSR scores

agents	proper	random	mean
trans-eurollm	0.37	0.57	0.47
trans-qwen3_8b	0.68	0.74	0.71
trans-qwen3_8b-think	0.72	0.74	0.73
trans-qwen3_14b	0.67	0.73	0.70
trans-qwen3_14b-think	0.70	0.77	0.73
trans-gemma3_12b	0.77	0.79	0.78
trans-gemma3_27b	0.78	0.77	0.78
revis-qwen3_14b-think	0.69	0.73	0.71
revis-gemma3_12b-think	0.72	0.76	0.74
revis-gemma3_27b-think	0.74	0.79	0.77
final	**0.86**	**0.87**	**0.87**

Table 26: TSR scores for Track1, translations from English to German.

agents	proper	random	mean
trans-eurollm	0.54	0.75	0.64
trans-qwen3_8b	0.78	0.81	0.79
trans-qwen3_8b-think	0.84	0.84	0.84
trans-qwen3_14b	0.78	0.81	0.79
trans-qwen3_14b-think	0.82	0.85	0.84
trans-gemma3_12b	0.83	0.85	0.84
trans-gemma3_27b	0.85	0.86	0.85
revis-qwen3_14b-think	0.80	0.80	0.80
revis-gemma3_12b-think	0.84	0.85	0.85
revis-gemma3_27b-think	0.85	0.86	0.85
final	**0.88**	**0.88**	**0.88**

Table 27: TSR scores for Track1, translations from English to Spanish.

agents	proper	random	mean
trans-eurollm	0.56	0.70	0.63
trans-qwen3_8b	0.80	0.84	0.82
trans-qwen3_8b-think	0.87	0.88	0.88
trans-qwen3_14b	0.83	0.84	0.84
trans-qwen3_14b-think	0.89	0.88	0.89
trans-gemma3_12b	0.89	0.88	0.89
trans-gemma3_27b	0.88	0.90	0.89
revis-qwen3_14b-think	0.84	0.84	0.84
revis-gemma3_12b-think	0.88	0.87	0.88
revis-gemma3_27b-think	0.88	0.87	0.88
final	**0.94**	**0.93**	**0.94**

Table 28: TSR scores for Track1, translations from English to Russian.

E.2 Track2

E.2.1 MetricX scores - English to Traditional Chinese

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	2.13	2.07	2.04	2.12	2.16	2.10
trans-qwen3_8b-think	2.07	2.04	1.98	2.02	2.05	2.03
trans-qwen3_14b	2.03	1.92	1.94	1.95	2.00	1.97
trans-qwen3_14b-think	1.97	1.90	1.83	1.89	1.90	1.90
trans-gemma3_12b	2.25	2.09	2.01	2.01	2.13	2.10
trans-gemma3_27b	2.05	1.92	1.89	1.99	2.01	1.97
revis-qwen3_14b-think	1.99	1.86	1.84	1.94	1.93	1.91
revis-gemma3_12b-think	1.99	1.93	1.83	1.88	1.96	1.92
revis-gemma3_27b-think	1.93	1.83	1.78	1.86	1.86	1.85
final	**1.49**	**1.40**	**1.37**	**1.41**	**1.46**	**1.43**

Table 29: MetricX scores for Track2, translations from English to Traditional Chinese with *no* terms.

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	2.43	2.38	2.22	2.36	2.41	2.36
trans-qwen3_8b-think	2.64	2.62	2.64	2.60	2.64	2.63
trans-qwen3_14b	**2.29**	**2.26**	**2.15**	**2.18**	**2.19**	**2.21**
trans-qwen3_14b-think	2.55	2.48	2.43	2.36	2.39	2.44
trans-gemma3_12b	2.96	2.85	2.70	2.66	2.84	2.80
trans-gemma3_27b	2.72	2.68	2.64	2.53	2.60	2.64
revis-qwen3_14b-think	2.39	2.40	2.37	2.39	2.42	2.40
revis-gemma3_12b-think	2.31	2.30	2.24	2.23	2.33	2.28
revis-gemma3_27b-think	2.46	2.36	2.41	2.29	2.35	2.38
final	2.39	2.28	2.26	2.23	2.22	2.28

Table 30: MetricX scores for Track2, translations from English to Traditional Chinese with *random* terms.

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	2.17	2.03	2.09	2.12	2.17	2.12
trans-qwen3_8b-think	2.19	2.04	2.09	2.09	2.14	2.11
trans-qwen3_14b	2.05	1.95	1.99	1.98	2.05	2.01
trans-qwen3_14b-think	2.05	1.90	1.92	1.97	2.02	1.97
trans-gemma3_12b	2.36	2.10	2.11	2.17	2.22	2.19
trans-gemma3_27b	2.18	1.98	2.01	2.04	2.10	2.06
revis-qwen3_14b-think	2.11	1.91	1.98	1.94	2.07	2.00
revis-gemma3_12b-think	2.09	1.93	2.04	2.00	2.01	2.01
revis-gemma3_27b-think	2.05	1.85	1.94	1.97	2.00	1.96
final	**1.88**	**1.70**	**1.73**	**1.77**	**1.84**	**1.78**

Table 31: `MetricX` scores for `Track2`, translations from English to Traditional Chinese with *proper* terms.

E.2.2 `MetricX` scores - Traditional Chinese to English - Bug

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.69	1.75	1.60	1.62	1.62	1.65
trans-qwen3_8b-think	1.64	1.66	1.61	1.59	1.62	1.62
trans-qwen3_14b	1.61	1.58	1.54	1.62	1.56	1.58
trans-qwen3_14b-think	1.59	1.57	1.51	1.57	1.52	1.55
trans-gemma3_12b	1.77	1.69	1.76	1.66	1.69	1.72
trans-gemma3_27b	1.66	1.66	1.62	1.62	1.65	1.64
revis-qwen3_14b-think	1.68	1.66	1.75	1.77	1.86	1.75
revis-gemma3_12b-think	1.63	1.66	1.65	1.67	1.70	1.66
revis-gemma3_27b-think	1.65	1.66	1.58	1.63	1.58	1.62
final	**1.29**	**1.26**	**1.18**	**1.22**	**1.20**	**1.23**

Table 32: `MetricX` scores for `Track2`, translations from Traditional Chinese to English with *random* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.76	1.74	1.60	1.68	1.69	1.69
trans-qwen3_8b-think	1.71	1.72	1.63	1.63	1.68	1.67
trans-qwen3_14b	1.62	1.63	1.58	1.63	1.68	1.63
trans-qwen3_14b-think	1.65	1.61	1.55	1.61	1.63	1.61
trans-gemma3_12b	1.75	1.76	1.65	1.72	1.75	1.73
trans-gemma3_27b	1.74	1.73	1.60	1.68	1.76	1.70
revis-qwen3_14b-think	1.83	1.71	1.80	1.98	1.92	1.85
revis-gemma3_12b-think	1.74	1.70	1.65	1.77	1.74	1.72
revis-gemma3_27b-think	1.72	1.66	1.62	1.64	1.68	1.66
final	**1.33**	**1.29**	**1.23**	**1.28**	**1.28**	**1.28**

Table 33: `MetricX` scores for `Track2`, translations from Traditional Chinese to English with *proper* terms.

E.2.3 `MetricX` scores - Traditional Chinese to English - Correct

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.69	1.69	1.57	1.59	1.60	1.63
trans-qwen3_8b-think	1.67	1.65	1.59	1.60	1.62	1.63
trans-qwen3_14b	1.57	1.57	1.52	1.62	1.56	1.57
trans-qwen3_14b-think	1.56	1.51	1.46	1.51	1.52	1.51
trans-gemma3_12b	1.59	1.54	1.51	1.47	1.51	1.52
trans-gemma3_27b	1.63	1.61	1.54	1.57	1.56	1.58
revis-qwen3_14b-think	2.19	2.13	2.08	2.07	2.07	2.11
revis-gemma3_12b-think	1.68	1.70	1.64	1.69	1.75	1.69
revis-gemma3_27b-think	1.67	1.62	1.59	1.61	1.60	1.62
final	**1.22**	**1.17**	**1.12**	**1.14**	**1.14**	**1.16**

Table 34: `MetricX` scores for `Track2`, translations from Traditional Chinese to English with *no* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.75	1.76	1.72	1.76	1.79	1.76
trans-qwen3_8b-think	1.77	1.96	1.86	1.73	1.85	1.83
trans-qwen3_14b	1.70	1.71	1.63	1.72	1.67	1.69
trans-qwen3_14b-think	1.74	1.79	1.75	1.67	1.70	1.73
trans-gemma3_12b	2.63	2.72	2.65	2.41	2.64	2.61
trans-gemma3_27b	2.09	2.37	2.10	1.96	2.03	2.11
revis-qwen3_14b-think	1.87	2.02	2.00	1.97	2.11	1.99
revis-gemma3_12b-think	1.87	2.09	1.92	1.84	1.92	1.93
revis-gemma3_27b-think	1.99	2.00	2.08	1.95	1.96	2.00
final	**1.54**	**1.58**	**1.53**	**1.46**	**1.52**	**1.53**

Table 35: `MetricX` scores for `Track2`, for translations from Traditional Chinese to English, with random terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	1.83	1.80	1.72	1.85	1.87	1.81
trans-qwen3_8b-think	1.91	1.88	1.83	1.94	1.90	1.89
trans-qwen3_14b	1.82	1.78	1.68	1.80	1.83	1.78
trans-qwen3_14b-think	1.88	1.86	1.72	1.84	1.85	1.83
trans-gemma3_12b	2.08	2.01	2.01	2.24	2.05	2.08
trans-gemma3_27b	2.10	2.05	1.94	2.10	2.02	2.04
revis-qwen3_14b-think	2.04	1.99	1.97	2.04	2.13	2.03
revis-gemma3_12b-think	1.99	1.92	1.90	2.01	1.93	1.95
revis-gemma3_27b-think	2.05	1.95	1.84	1.97	1.91	1.95
final	**1.59**	**1.57**	**1.47**	**1.58**	**1.53**	**1.55**

Table 36: `MetricX` scores for `Track2`, for translations from Traditional Chinese to English, with proper terms.

E.2.4 `MetricX` scores - Traditional Chinese to English - Bug Assessment

We are unable to adequately assess the expected `MetricX` scores and the drop in the final results of our submission as a result of the bug. For this assessment, we need access to the translation reference texts that are unavailable. However, as we can observe in Sections E.1.1, E.2.1, E.2.2 and E.2.3, `MetricX` scores are increasing (i.e. translation quality drops) for random and proper term cases compared to the

no-term case. The intuition behind this trend is that forcing terminology reduces the general translation quality because translators need to balance between two different objectives: 1) general translation and 2) terminology constraint. Based on this intuition, we can approximate the drop in the final MetricX scores as a missed increase in MetricX scores (i.e. missed drop in general translation quality) as a result of less constrained terminology.

agents	2016	2018	2020	2022	2024	mean
final-bug	1.29	1.26	1.18	1.22	1.20	1.23
final-correct	1.54	1.58	1.53	1.46	1.52	1.53
final-diff	0.25	0.32	0.35	0.24	0.32	0.30

Table 37: Estimation of the final score drop due to the bug, MetricX scores for Track2, translations from Traditional Chinese to English with *random* terms.

agents	2016	2018	2020	2022	2024	mean
final-bug	1.33	1.29	1.23	1.28	1.28	1.28
final-correct	1.59	1.57	1.47	1.58	1.53	1.55
final-diff	0.26	0.28	0.24	0.30	0.25	0.27

Table 38: Estimation of the final score drop due to the bug, MetricX scores for Track2, translations from Traditional Chinese to English with *proper* terms.

E.2.5 TSR scores - English to Traditional Chinese

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	0.56	0.47	0.50	0.53	0.58	0.53
trans-qwen3_8b-think	0.55	0.51	0.55	0.58	0.58	0.55
trans-qwen3_14b	0.53	0.42	0.49	0.47	0.52	0.49
trans-qwen3_14b-think	0.53	0.48	0.51	0.50	0.51	0.51
trans-gemma3_12b	0.61	0.54	0.58	0.56	0.60	0.58
trans-gemma3_27b	0.60	0.54	0.60	0.59	0.65	0.60
revis-qwen3_14b-think	0.53	0.41	0.50	0.48	0.54	0.49
revis-gemma3_12b-think	0.55	0.44	0.53	0.51	0.57	0.52
revis-gemma3_27b-think	0.58	0.47	0.56	0.52	0.59	0.55
final	**0.75**	**0.70**	**0.72**	**0.73**	**0.75**	**0.73**

Table 39: TSR scores for Track2, for translations from English to Traditional Chinese, with *random* terms, Track2.

agents	2015	2017	2019	2021	2023	mean
trans-qwen3_8b	0.65	0.64	0.64	0.66	0.67	0.65
trans-qwen3_8b-think	0.66	0.65	0.66	0.67	0.68	0.66
trans-qwen3_14b	0.66	0.64	0.64	0.66	0.67	0.66
trans-qwen3_14b-think	0.66	0.65	0.66	0.68	0.68	0.67
trans-gemma3_12b	0.70	0.69	0.68	0.72	0.71	0.70
trans-gemma3_27b	0.70	0.69	0.69	0.72	0.71	0.70
revis-qwen3_14b-think	0.67	0.66	0.67	0.68	0.68	0.67
revis-gemma3_12b-think	0.67	0.67	0.67	0.69	0.68	0.68
revis-gemma3_27b-think	0.68	0.68	0.67	0.69	0.69	0.68
final	**0.78**	**0.77**	**0.78**	**0.81**	**0.79**	**0.79**

Table 40: TSR scores for Track2, for translations from English to Traditional Chinese, with *proper* terms, Track2.

E.2.6 TSR scores - Traditional Chinese to English - Bug

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.56	0.90	0.90	0.90	0.86	0.82
trans-qwen3_8b-think	0.55	0.90	0.90	0.90	0.86	0.82
trans-qwen3_14b	0.54	0.90	0.90	0.90	0.86	0.82
trans-qwen3_14b-think	0.55	0.90	0.90	0.90	0.86	0.82
trans-gemma3_12b	0.55	0.89	0.90	0.90	0.86	0.82
trans-gemma3_27b	0.55	0.89	0.90	0.91	0.86	0.82
revis-qwen3_14b-think	0.55	0.89	0.90	0.90	0.85	0.82
revis-gemma3_12b-think	0.55	0.89	0.90	0.90	0.85	0.82
revis-gemma3_27b-think	0.55	0.89	0.89	0.90	0.86	0.82
final	**0.56**	**0.91**	**0.90**	**0.91**	**0.87**	**0.83**

Table 41: TSR scores for Track2, for translations from Traditional Chinese to English, with *random* terms, Track2.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.90	0.88	0.91	0.89	0.90	0.89
trans-qwen3_8b-think	0.91	0.88	0.91	0.89	0.89	0.90
trans-qwen3_14b	0.90	0.89	0.91	0.90	0.90	0.90
trans-qwen3_14b-think	0.91	0.90	0.92	0.91	0.90	0.91
trans-gemma3_12b	0.91	0.87	0.91	0.88	0.89	0.89
trans-gemma3_27b	0.91	0.88	0.90	0.88	0.89	0.89
revis-qwen3_14b-think	0.91	0.90	0.91	0.89	0.90	0.90
revis-gemma3_12b-think	0.90	0.87	0.90	0.90	0.89	0.89
revis-gemma3_27b-think	0.91	0.88	0.91	0.90	0.90	0.90
final	**0.92**	**0.91**	**0.92**	**0.92**	**0.91**	**0.92**

Table 42: TSR scores for Track2, for translations from Traditional Chinese to English, with *proper* terms, Track2.

E.2.7 TSR scores - Traditional Chinese to English - Correct

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.35	0.43	0.44	0.44	0.43	0.42
trans-qwen3_8b-think	0.36	0.43	0.44	0.44	0.42	0.42
trans-qwen3_14b	0.34	0.43	0.44	0.44	0.42	0.42
trans-qwen3_14b-think	0.35	0.43	0.44	0.44	0.42	0.42
trans-gemma3_12b	0.35	0.43	0.44	0.45	0.44	0.42
trans-gemma3_27b	0.36	0.44	0.45	0.45	0.43	0.43
revis-qwen3_14b-think	0.34	0.42	0.43	0.43	0.41	0.41
revis-gemma3_12b-think	0.36	0.42	0.43	0.44	0.42	0.41
revis-gemma3_27b-think	0.35	0.44	0.45	0.45	0.43	0.42
final	**0.39**	**0.48**	**0.48**	**0.48**	**0.46**	**0.46**

Table 43: TSR scores for Track2, for translations from Traditional Chinese to English, with *random* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.66	0.70	0.73	0.76	0.71	0.71
trans-qwen3_8b-think	0.66	0.70	0.73	0.75	0.71	0.71
trans-qwen3_14b	0.66	0.71	0.73	0.77	0.72	0.72
trans-qwen3_14b-think	0.67	0.72	0.73	0.77	0.72	0.72
trans-gemma3_12b	0.66	0.69	0.72	0.75	0.70	0.70
trans-gemma3_27b	0.66	0.70	0.72	0.76	0.72	0.71
revis-qwen3_14b-think	0.67	0.71	0.72	0.76	0.72	0.72
revis-gemma3_12b-think	0.66	0.68	0.71	0.76	0.71	0.70
revis-gemma3_27b-think	0.66	0.70	0.73	0.76	0.72	0.71
final	**0.69**	**0.74**	**0.75**	**0.80**	**0.75**	**0.75**

Table 44: TSR scores for Track2, for translations from Traditional Chinese to English, with *proper* terms.

E.2.8 TSR scores - Traditional Chinese to English - Bug Assessment

Here, we provide an assessment of the expected TSR scores and the drop in the final results of our submission as a result of the bug. The assessment is calculated with full correct terms on the original submitted translations that contain the bug.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.34	0.41	0.41	0.42	0.40	0.40
trans-qwen3_8b-think	0.33	0.42	0.41	0.42	0.40	0.40
trans-qwen3_14b	0.33	0.41	0.41	0.42	0.40	0.39
trans-qwen3_14b-think	0.33	0.41	0.41	0.42	0.40	0.39
trans-gemma3_12b	0.33	0.40	0.41	0.42	0.39	0.39
trans-gemma3_27b	0.33	0.40	0.41	0.42	0.40	0.39
revis-qwen3_14b-think	0.33	0.40	0.40	0.41	0.39	0.39
revis-gemma3_12b-think	0.33	0.40	0.40	0.41	0.39	0.38
revis-gemma3_27b-think	0.33	0.41	0.40	0.42	0.40	0.39
final	**0.36**	**0.45**	**0.44**	**0.45**	**0.43**	**0.43**
final-delta (bug - correct)	**-0.03**	**-0.03**	**-0.04**	**-0.03**	**-0.03**	**-0.03**

Table 45: Assessment of the final score drop due to the bug, TSR scores for Track2, for translations from Traditional Chinese to English, with *random* terms.

agents	2016	2018	2020	2022	2024	mean
trans-qwen3_8b	0.53	0.53	0.54	0.61	0.55	0.55
trans-qwen3_8b-think	0.54	0.54	0.56	0.61	0.56	0.56
trans-qwen3_14b	0.56	0.59	0.57	0.65	0.59	0.59
trans-qwen3_14b-think	0.56	0.59	0.59	0.65	0.59	0.60
trans-gemma3_12b	0.52	0.53	0.54	0.60	0.54	0.55
trans-gemma3_27b	0.55	0.56	0.56	0.64	0.58	0.58
revis-qwen3_14b-think	0.57	0.58	0.58	0.64	0.59	0.59
revis-gemma3_12b-think	0.55	0.55	0.56	0.63	0.56	0.57
revis-gemma3_27b-think	0.55	0.56	0.57	0.64	0.57	0.58
final	**0.61**	**0.63**	**0.64**	**0.69**	**0.64**	**0.64**
final-delta (bug - correct)	**-0.08**	**-0.11**	**-0.11**	**-0.11**	**-0.11**	**-0.11**

Table 46: Assessment of the final score drop due to the bug, TSR scores for Track2, for translations from Traditional Chinese to English, with *proper* terms.

E.3 Metric-guided agent selection

E.3.1 Track1

agents	noterm	proper	random	total
trans-eurollm	10.4%	6.2%	12.6%	9.7%
trans-qwen3_8b	4.6%	7.0%	4.8%	5.4%
trans-qwen3_8b-think	5.2%	5.2%	5.2%	5.2%
trans-qwen3_14b	6.4%	5.0%	7.4%	6.2%
trans-qwen3_14b-think	3.6%	5.4%	5.6%	4.8%
trans-gemma3_12b	8.4%	9.2%	8.2%	8.6%
trans-gemma3_27b	7.2%	7.4%	5.8%	6.8%
revis-qwen3_14b-think	12.0%	12.4%	10.2%	11.5%
revis-gemma3_12b-think	**23.5%**	**24.6%**	**26.4%**	**24.8%**
revis-gemma3_27b-think	18.6%	17.5%	13.8%	16.6%

Table 47: The frequency of agents selected for the final solution, for translations from English to German, Track1.

agents	noterm	proper	random	total
trans-eurollm	12.2%	13.2%	11.7%	12.4%
trans-qwen3_8b	3.4%	6.4%	5.6%	5.1%
trans-qwen3_8b-think	8.2%	5.0%	5.8%	6.3%
trans-qwen3_14b	6.2%	8.7%	6.0%	7.0%
trans-qwen3_14b-think	4.0%	5.0%	3.2%	4.0%
trans-gemma3_12b	9.6%	4.3%	5.6%	6.5%
trans-gemma3_27b	8.0%	7.0%	7.0%	7.3%
revis-qwen3_14b-think	13.0%	7.6%	9.4%	10.0%
revis-gemma3_12b-think	17.2%	**27.6%**	**31.2%**	**25.3%**
revis-gemma3_27b-think	**18.2%**	15.0%	14.4%	15.8%

Table 48: The frequency of agents selected for the final solution, for translations from English to Spanish, Track1.

agents	noterm	proper	random	total
trans-eurollm	11.6%	9.6%	11.7%	11.0%
trans-qwen3_8b	4.3%	5.2%	5.2%	4.9%
trans-qwen3_8b-think	5.4%	6.2%	7.0%	6.2%
trans-qwen3_14b	7.2%	6.0%	5.0%	6.0%
trans-qwen3_14b-think	3.2%	4.6%	4.2%	4.0%
trans-gemma3_12b	10.6%	8.6%	10.2%	9.8%
trans-gemma3_27b	7.4%	7.0%	7.8%	7.4%
revis-qwen3_14b-think	11.0%	10.8%	10.2%	10.6%
revis-gemma3_12b-think	**21.4%**	**27.8%**	**25.2%**	**24.8%**
revis-gemma3_27b-think	17.8%	14.2%	13.4%	15.1%

Table 49: The frequency of agents selected for the final solution, for translations from English to Russian, Track1.

E.3.2 Track2

agents	2015	2017	2019	2021	2023	total
trans-qwen3_8b	11.0%	6.4%	6.9%	7.4%	6.0%	7.4%
trans-qwen3_8b-think	7.8%	4.8%	5.1%	6.1%	6.6%	6.1%
trans-qwen3_14b	6.1%	6.7%	6.3%	7.9%	6.8%	6.8%
trans-qwen3_14b-think	7.5%	7.1%	10.1%	9.0%	9.4%	8.7%
trans-gemma3_12b	7.2%	11.1%	12.1%	12.0%	13.8%	11.6%
trans-gemma3_27b	9.3%	11.8%	13.5%	8.8%	10.1%	10.6%
revis-qwen3_14b-think	12.2%	14.6%	14.2%	12.3%	10.5%	12.6%
revis-gemma3_12b-think	17.7%	**19.0%**	**16.2%**	17.3%	16.6%	17.3%
revis-gemma3_27b-think	**20.6%**	18.0%	15.1%	**18.7%**	**19.7%**	**18.4%**

Table 50: The frequency of agents selected for the final solution, for translations from English to Traditional Chinese, with no terms, Track2.

agents	2015	2017	2019	2021	2023	total
trans-qwen3_8b	7.2%	7.6%	6.0%	4.7%	5.3%	6.0%
trans-qwen3_8b-think	5.2%	6.9%	5.6%	6.5%	7.3%	6.4%
trans-qwen3_14b	6.4%	6.7%	8.3%	6.9%	5.9%	6.8%
trans-qwen3_14b-think	8.7%	6.2%	9.2%	10.0%	6.6%	8.2%
trans-gemma3_12b	13.1%	15.5%	12.4%	15.0%	15.5%	14.4%
trans-gemma3_27b	12.8%	9.9%	12.4%	9.7%	14.0%	11.7%
revis-qwen3_14b-think	9.6%	11.3%	11.2%	13.4%	12.1%	11.7%
revis-gemma3_12b-think	**20.6%**	**19.4%**	**19.4%**	**20.1%**	**20.5%**	**20.0%**
revis-gemma3_27b-think	16.0%	16.0%	15.1%	13.2%	12.3%	14.3%

Table 51: The frequency of agents selected for the final solution, for translations from English to Traditional Chinese, with *proper* terms, Track2.

agents	2015	2017	2019	2021	2023	total
trans-qwen3_8b	9.0%	11.6%	9.7%	7.4%	6.4%	8.6%
trans-qwen3_8b-think	6.1%	5.5%	8.5%	7.4%	8.3%	7.3%
trans-qwen3_14b	9.3%	6.2%	6.5%	6.3%	7.3%	7.0%
trans-qwen3_14b-think	5.2%	7.1%	8.5%	7.4%	6.6%	7.1%
trans-gemma3_12b	15.4%	**17.6%**	**15.1%**	**18.0%**	13.1%	15.8%
trans-gemma3_27b	15.7%	16.0%	14.8%	14.8%	15.7%	15.4%
revis-qwen3_14b-think	8.1%	7.6%	9.2%	6.7%	9.6%	8.2%
revis-gemma3_12b-think	**17.4%**	12.7%	14.8%	16.8%	**18.2%**	**16.1%**
revis-gemma3_27b-think	13.4%	15.3%	12.4%	14.8%	14.4%	14.1%

Table 52: The frequency of agents selected for the final solution, for translations from English to Traditional Chinese, with *random* terms, Track2.

agents	2016	2018	2020	2022	2024	total
trans-qwen3_8b	7.0%	8.6%	8.5%	8.6%	9.8%	8.6%
trans-qwen3_8b-think	8.6%	5.0%	7.2%	7.5%	7.2%	7.1%
trans-qwen3_14b	9.4%	8.4%	10.9%	8.6%	9.4%	9.4%
trans-qwen3_14b-think	10.0%	10.3%	9.3%	9.6%	8.7%	9.5%
trans-gemma3_12b	13.2%	13.7%	**15.5%**	**17.3%**	**15.3%**	15.2%
trans-gemma3_27b	9.1%	9.3%	9.3%	9.4%	11.0%	9.7%
revis-qwen3_14b-think	15.1%	**18.2%**	14.6%	16.3%	**15.3%**	**15.8%**
revis-gemma3_12b-think	**17.2%**	13.7%	13.4%	13.6%	11.2%	13.5%
revis-gemma3_27b-think	10.0%	12.5%	10.9%	8.6%	11.7%	10.7%

Table 53: The frequency of agents selected for the final solution, for translations from Traditional Chinese to English, with *no* terms, Track2.

agents	2016	2018	2020	2022	2024	total
trans-qwen3_8b	11.0%	8.8%	8.9%	8.4%	11.3%	9.7%
trans-qwen3_8b-think	7.0%	5.2%	5.0%	6.7%	6.3%	6.0%
trans-qwen3_14b	10.2%	11.5%	13.6%	10.7%	10.5%	11.3%
trans-qwen3_14b-think	9.1%	8.1%	6.8%	10.0%	8.7%	8.5%
trans-gemma3_12b	10.5%	10.8%	9.7%	9.6%	9.1%	9.8%
trans-gemma3_27b	8.9%	10.3%	11.6%	10.4%	11.0%	10.5%
revis-qwen3_14b-think	15.1%	**19.4%**	**19.2%**	**17.7%**	**15.6%**	**17.4%**
revis-gemma3_12b-think	**15.6%**	16.1%	13.6%	15.7%	13.6%	14.8%
revis-gemma3_27b-think	12.1%	9.3%	11.1%	10.4%	13.4%	11.3%

Table 54: The frequency of agents selected for the final solution, for translations from Traditional Chinese to English, with *proper* terms, Track2.

agents	2016	2018	2020	2022	2024	total
trans-qwen3_8b	7.2%	6.9%	7.6%	8.2%	9.3%	8.0%
trans-qwen3_8b-think	6.7%	8.1%	7.0%	7.7%	7.0%	7.3%
trans-qwen3_14b	8.1%	10.5%	9.5%	9.8%	14.3%	10.7%
trans-qwen3_14b-think	7.8%	10.0%	6.4%	8.4%	7.5%	8.0%
trans-gemma3_12b	14.0%	10.5%	12.8%	11.9%	13.6%	12.6%
trans-gemma3_27b	11.6%	12.7%	11.1%	13.1%	9.3%	11.4%
revis-qwen3_14b-think	**16.4%**	**17.3%**	**18.5%**	**14.4%**	**14.3%**	**16.0%**
revis-gemma3_12b-think	16.2%	11.0%	12.2%	12.9%	11.5%	12.6%
revis-gemma3_27b-think	11.6%	12.5%	14.6%	13.2%	12.9%	13.0%

Table 55: The frequency of agents selected for the final solution, for translations from Traditional Chinese to English, with *random* terms, Track2.

E.4 Final results - Track2- Bug Correction

Here we compare the Track2 scores of the submitted MeGuMa system and the scores of the error-corrected version. The error occurred when projecting document-level terminology map to the paragraph level. This operation, performed because our system operates on the paragraph level, used whitespace-delimiters for term matching. This approach, suitable for European languages, is not correct for Chinese texts. For a more detailed description, see the end of Subsection 2.

The corrected system was scored using the evaluation code from the official repository.[5] The corrected show a large increase (34 points) of the terminology success rate Term-Acc (labeled as "Proper, Acc." by the organizers). This is expected, since the erroneous term matching caused the loss of predefined term translations fed to the system. For the random terminology, there is a smaller increase of 2 points. Translation accuracies, in terms of ChrF2++, increase by approximately 3.5 points.

In order to ensure transparency, our repository[6] contains detailed code-level description of the bug, the outputs of the corrected system, and the instructions how to run our system.

System	Bleu4	ChrF	Proper, Acc.	Random, Acc.
MeGuMa [submitted]	32.96	69.41	62.43	86.76
MeGuMa [debugged]	39.07	72.73	96.62	87.66
Difference	6.11	3.32	34.19	0.90

Table 56: The final scores of MeGuMa system, before and after correcting the data preparation bug. Translations from Traditional Chinese to English, with *proper* terms, Track2.

System	Bleu4	ChrF	Proper, Acc.	Random, Acc.
MeGuMa [submitted]	23.88	65.21	51.55	86.44
MeGuMa [debugged]	31.38	69.07	53.68	92.58
Difference	7.50	3.86	2.13	6.14

Table 57: The final scores of MeGuMa system, before and after correcting the data preparation bug. Translations from Traditional Chinese to English, with *random* terms, Track2.

System	Bleu4	ChrF	Proper, Acc.	Random, Acc.
MeGuMa [submitted]	30.83	68.31	51.91	85.91
MeGuMa [debugged]	30.83	68.31	51.91	85.91
Difference	0.00	0.00	0.00	0.00

Table 58: The final scores of MeGuMa system, before and after correcting the data preparation bug. Translations from Traditional Chinese to English, with *noterm* terms, Track2.

[5] https://github.com/wmt-conference/wmt25-terminology/
[6] https://github.com/igrubi/irb-mt-wmt2025

Terminology-Constrained Translation from Monolingual Data using GRPO

Javier Garcia Gilabert[1] **Carlos Escolano**[1,2] **Xixian Liao**[1] **Maite Melero**[1]
[1]Barcelona Supercomputing Center
[2]Universitat Politècnica de Catalunya

Abstract

Terminology consistency is essential for high-quality machine translation, especially in domain-specific and professional contexts, where accurate term translation directly impacts usability. This paper presents the submission from the BSC team to the WMT25 Terminology-Aware Translation Task. We propose the use of GRPO (Group Relative Policy Optimization) to adapt translation models using monolingual data only, without requiring parallel corpora. Our reward function jointly optimizes for terminology adherence and overall translation quality, leveraging quality-estimation metrics. Experimental results demonstrate that our method consistently improves terminology translation across three language directions—English to Spanish, German, and Russian—by up to +0.36 T_P points across all evaluated models.

1 Introduction

In recent years, large language models (LLMs) have emerged as the state of the art across a wide range of natural language processing tasks, including machine translation (MT). Owing to their vast number of parameters, these models possess the capacity to handle multiple languages and adapt across diverse domains. However, their large scale also makes them computationally expensive to fine-tune or adapt to specific settings.

In this work, we address the task of terminology-aware machine translation, aiming to produce domain-specific translations without the need for costly fine-tuning. Rather than adapting the model through retraining, we condition it at inference time by incorporating a bilingual glossary into the translation prompt. This approach enables flexible, on-the-fly domain adaptation, assuming that relevant terminology is available.

Another major challenge in adapting LLMs to MT tasks lies in the scarcity of high-quality parallel corpora, particularly for low-resource language pairs. In many cases, obtaining sufficient parallel data for supervised training is infeasible.

Prompt: Translate the following text from English into Spanish. Please ensure the following terminology is used:

- advertising campaigns → acciones publicitarias
- social platforms → redes sociales

English: "Spotlight's investigation notes that other countries also ran advertising campaigns, including Malta, Greece, Albania, Poland, Armenia and France. Some of these were run by the artists themselves via their own social platforms."

Spanish: La investigación de Spotlight señala que otros países también llevaron a cabo acciones publicitarias –entre ellos, Malta, Grecia, Albania, Polonia, Armenia y Francia–. Algunas de estas campañas fueron organizadas por los propios artistas a través de sus propias redes sociales.

Figure 1: Example of prompt for terminology-aware translations with a glossary. In green, source terms in English. In yellow, target translations in Spanish.

This paper presents the Barcelona Supercomputing Center (BSC) submission to the terminology-aware MT task for the first track: *Sentence/Paragraph-Level Translation*. Our system is built around two key contributions:

- Application of Group Relative Policy Optimization (GRPO) (DeepSeek-AI et al., 2025; Shao et al., 2024): We demonstrate how GRPO can be used to efficiently instruct LLMs in terminology-aware translation. Our experiments show significant performance improvements across multiple language pairs and model architectures.
- Leveraging monolingual data with quality estimation: We show that incorporating quality estimation metrics, such as COMET-KIWI (Rei et al., 2022), allows the model to benefit from

Proceedings of the Tenth Conference on Machine Translation, pages 1335–1343
November 8-9, 2025

monolingual data alone, eliminating the need for parallel corpora in supervised training.

2 Related Work

Prior work in terminology-aware machine translation has taken several different approaches. A common strategy involves fine-tuning models with terminology constraints. For instance, Kim et al. (2024) extract terminology from training data to build a glossary and fine-tune the model using inputs augmented with extracted terms. Zheng et al. (2024) propose DragFT, a framework combining dictionary-enhanced prompting, retrieval-augmented few-shot selection, and fine-tuning to improve translation in specialized domains. Another line of work focuses on synthetic data generation and post-editing. Moslem et al. (2023) use LLMs to generate bilingual data containing pre-specified terminology, which is then used to fine-tune MT models. They further apply LLM-based post-editing to insert missing terms into system outputs that failed to adhere to terminology constraints. Other methods aim to enforce terminology during decoding. Bogoychev and Chen (2023) explore constrained decoding strategies and LLM-based paraphrasing to increase term fidelity, including the use of negative constraints that penalize incorrect term usage. Reinforcement learning (RL) has also been explored as a way to improve terminology translation. Li et al. (2025) integrate RL with word alignment to define reward signals, enabling models to translate key terminology without explicit term detection at inference time. Our work is most closely aligned with this last line of research. The key distinction is that our approach relies solely on monolingual data and inference-time prompting, and is therefore better suited to low-resource settings where parallel corpora may be scarce or unavailable.

3 Methodology

In this section, we will discuss our proposed method to adapt the LLMs to the task of terminology-aware translation, using monolingual data only.

3.1 Data Preparation

Given that our training data is monolingual, we first create a bilingual glossary containing some of the terms from our source text. To do so, given an English source text, we employ the `spaCy` library (Honnibal et al., 2020) to identify candidate terminology phrases. The extraction heuristic combines three types of linguistic units:

- **Named Entities** Matches proper nouns and numerical expressions (e.g., organizations, locations, dates).
- **Noun Phrases** Captures syntactic chunks that often represent key domain-specific concepts.
- **Adverbial Constructions** Extracts adverbs modifying verbs, adjectives, or other adverbs to capture domain-relevant descriptions.

From these candidates, we select up to five non-overlapping phrases per text. Then, each extracted term is individually translated into the target language using the NLLB 3.3B (Costa-jussà et al., 2022) machine translation model. Once the pairs are generated, all examples are formatted as prompts following the template shown in Appendix Figure 4. Figure 1 shows an example for the English-Spanish pair.

3.2 Group Relative Policy Optimization

In order to adapt the LLMs, we employ Group Relative Policy Optimization (GRPO). This technique allows for efficient training using reinforcement learning without the need for an additional critic model (Schulman et al., 2017; Rafailov et al., 2023). In each training step, for each source sentence q, we sample G candidate translations $\{o_1, o_2, \cdots, o_G\}$ from the current policy model π. Then, we optimize the model parameters maximizing the following objective:

$$\frac{1}{G}\sum_{1}^{G}(\min(\nabla_\pi A_i, \text{clip}(\nabla_\pi, 1-\epsilon, 1+\epsilon)A_i) - \beta\mathcal{D} \quad (1)$$

$$\nabla_\pi = \frac{\pi'(o_i|q)}{\pi_{ref}(o_i|q)} \quad (2)$$

where π' is the adapted model and π_{ref} is the original model used for regularization. The objective function is composed of two terms. The first one computes the average of the losses for all outputs o_i generated from source sentence q. Each output's loss is defined as the minimum of the clipped and unclipped division of the output probabilities of the adapted model by the output probabilities of the original model multiplied by the advantage A_i. Finally, the second term of the loss

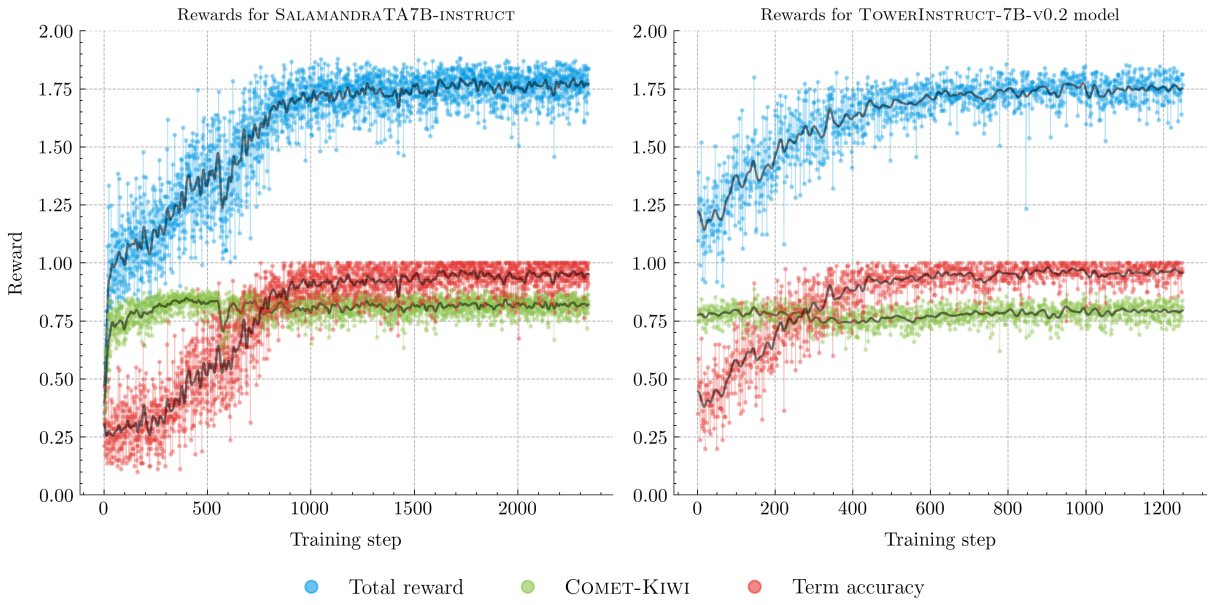


Figure 2: Reward evolution during training for SalamandraTA7B-Instruct (Left) and TowerInstruct-7B-v.02 (Right).

is the Kullback–Leibler distance ($\mathbb{D}_{kl}$) between the output distribution of the adapted model and the original model which is computed as follows:

$$\mathcal{D} = \mathbb{D}_{kl}(\pi' || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi'(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi'(o_i|q)} - 1 \quad (3)$$

Note that there are two hyperparameters that need to be set. First, ϵ controls the PPO clipping threshold. Second, β controls the Kullback–Leibler penalty which measures the relative entropy between both distributions. This penalty prevents the adapted model from diverging too far from the original model, which could cause performance degradation.

The advantage is computed as the normalized reward r_i from each output translation o_i as follows:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \quad (4)$$

where r_i denotes the overall reward of o_i. We define our reward function as a sum of two terms: (1) a terminology adherence score that evaluates the correct usage of terms from a provided glossary given in the prompt, and (2) a translation faithfulness score derived from an automated Quality Estimation (QE) metric. This last term is intended to regularize the training process, preventing the model from sacrificing overall translation quality in order to maximize terminology adherence, a behavior that can be seen as reward hacking. The final reward r_i is a linear combination of these two scores:

$$r_i = S_i + \gamma(o_i, q) \quad (5)$$

where S_i denotes the terminology adherence score while $\gamma(o_i, q)$ measures the translation faithfulness of the candidate translation (o_i) with respect to source sentence (q). In this work, we experiment with COMET-Kiwi[1] (Rei et al., 2022) as the quality estimation metric.

Terminology adherence reward The aim of this metric is to compute the proportion of terms in the glossary that are included in the candidate translation. We define the adherence score S_i of a candidate translation o_i as follows:

$$S_i = \frac{1}{|T|}\sum_{i=1}^{|T|}\delta(t_i \in o_i) \quad (6)$$

where T is a glossary of bilingual terms, $|T|$ is the number of terms, $t_i \in T$ is each individual term in the glossary and δ is a transformation of the term to adjust to the translation. For these experiments it will be set to the Identity, but it could be, for example, lemmatization or stemming, allowing the metric to account for morphological variations of the terms (e.g., "run" vs "running"). The score ranges from 0 (no adherence) to 1 (full adherence).

Translation faithfulness reward Despite the adherence score ensuring that the produced translations include the terminology, there are two additional aspects to consider. First, previous work on

[1] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

Direction	Model	T_P	T_F	BLEU	CHRF	COMET
En → Es	TowerInstruct-7B-v0.2	0.48	0.49	27.43	45.83	0.74
	+ GRPO	**0.93**	**0.91**	**51.27**	**74.21**	0.89
	SalamandraTA7B-instruct	0.54	0.54	43.64	62.82	0.79
	+ GRPO	0.90	0.88	47.46	73.75	**0.90**
En → De	TowerInstruct-7B-v0.2	0.60	0.59	38.81	65.43	0.86
	+ GRPO	**0.90**	**0.88**	39.40	68.33	0.87
	SalamandraTA7B-instruct	0.66	0.66	24.57	46.09	0.70
	+ GRPO	0.89	0.87	**44.46**	**71.26**	**0.89**
En → Ru	TowerInstruct-7B-v0.2	0.54	0.57	27.64	58.90	0.87
	+ GRPO	**0.87**	**0.86**	26.08	60.58	0.85
	SalamandraTA7B-instruct	0.66	0.68	20.70	45.10	0.72
	+ GRPO	0.84	0.85	**30.91**	**63.17**	**0.88**

Table 1: Performance of TowerInstruct-7B-v0.2 and SalamandraTA7B-instruct models on terminology-aware translation for English-to-Spanish (En→Es), English-to-German (En→De), and English-to-Russian (En→Ru) directions. Results are reported for both base models and models aligned with GRPO.

using GRPO for Machine Translation (Feng et al., 2025) has observed reward hacking, a phenomenon where models trained on a reward may produce answers that satisfy the reward but fail to solve the tasks. In the terminology task, an example would be copying the glossary without producing a translation. The second concern is that the proposed reward does not take translation quality into consideration, which may lead to catastrophic forgetting of translation quality during training. To prevent these behaviors, we introduce a second term for the reward, where we optimize COMET-KIWI, a quality estimation metric that allows us to evaluate translation quality by computing the similarity between the source sentence and the translation, without requiring a reference translation. As with the terminology adherence score, the faithfulness reward ranges from 0 to 1, with values closer to 1 indicating higher faithfulness.

4 Experiments

4.1 Models

Our experiments required LLMs with strong performance in machine translation. For this reason, we chose two models that were specifically adapted to this task, built on top of generalist LLMs:

TowerInstruct-7B-v.02 (Rei et al., 2024) This model is an adaptation of Llama2-7B for the task of machine translation across ten different languages (English, Portuguese, French, German, Russian, Chinese, Spanish, Dutch, Korean, and Italian). TowerInstruct-7B-v.02 was trained in two main steps: (1) continual pre-training on a combination of monolingual and parallel data; (2) instruction tuning on various tasks such as named entity recognition, machine translation, and post-editing.

SalamandraTA7B-Instruct (Gilabert et al., 2025) This model is an adaptation of Salamandra-7B (Gonzalez-Agirre et al., 2025) for machine translation. It supports 35 languages, including all the official European languages plus several regional Spanish languages such as Catalan, Basque, Galician, and Aranese. The model follows a similar approach to Tower LLM (Alves et al., 2024), with a continual pre-training phase over 424 billion tokens across all supported languages pivoting over Catalan, Spanish, and English. This is followed by an instruction tuning phase on tasks such as paragraph-level translation, post-editing, and alternative translations.

4.2 Evaluation

We evaluate our trained models on two aspects: terminology accuracy and general translation quality. To measure terminology accuracy, we use a provided glossary to compute: (1) Terminology Precision (T_P), the proportion of correctly translated terms, computed using an exact regular-expression match against the reference; and (2) Fuzzy Terminology Precision (T_F), which uses fuzzy matching with an 80% similarity threshold to account

for minor orthographic variations[2]. To assess general translation quality, we evaluate models before and after adaptation using the n-gram-based metrics BLEU[3] (Papineni et al., 2002) and CHRF[4] (Popović, 2015), and the embedding-based metric COMET[5] (Rei et al., 2020) for translation quality.

4.3 Data and Implementation

All experiments were conducted on a subset of the English portion of the *News-Commentary* dataset (Tiedemann, 2012). Model performance was evaluated on the development set released for the WMT25 shared task using the proper terminology subset.

Training was performed with a learning rate of 5×10^{-7} and a temperature parameter of 1.0 for sampling. The maximum generation length was limited to 1024 tokens. To prevent exploding gradients, we applied gradient clipping with a maximum norm of 1.0. Following Feng et al. (2025), we set the GRPO β hyperparameter to 0, and ϵ to 0.3 during training. All models were trained using the `Verl` framework (Sheng et al., 2024) for reinforcement learning, on four NVIDIA H100 GPUs with 64GB RAM each.

4.4 Experimental Results

During training, our first concern was to ensure that the proposed rewards provided enough signal for the model to adapt to the tasks. Figure 2 shows the variation of the two terms of the reward during the training process. We observe that the terminology adherence reward significantly increases during the first training updates, rising from approximately 0.25 accuracy to nearly 1, indicating that, by the end of training, the proposed translations included the terminology in almost all cases.

When looking at the COMET-KIWI score, we observe some differences between the two models. In TowerInstruct-7B-v.02, this reward remains constant throughout training, while SalamandraTA7B-Instruct shows an increase during the first updates of the training. This behavior may be related to greater improvements in translation performance for the latter model. After this initial increase, the COMET-KIWI reward stabilizes.

From these results, we observe that the COMET-KIWI reward functions as a regularization score, maintaining overall translation quality, while the terminology-adherence score is primarily responsible for guiding the model to produce the terminology defined by the glossary.

When looking at the model results in Table 1, we draw similar conclusions. Both models show significant performance gains in terminology accuracy across the directions tested. TowerInstruct-7B-v.02 achieves an average improvement of 0.36 T_P, while SalamandraTA7B-Instruct shows an average improvement of 0.29 T_P.

When looking at the translation performance, we observe more differences between the two models. TowerInstruct-7B-v.02 appears to exhibit inconsistent behavior across languages. While English to Spanish shows significant gains (over 20 BLEU points or 0.15 COMET), English to German shows only small improvements, and in the case of English to Russian even a performance degradation.

Meanwhile, SalamandraTA7B-Instruct shows consistent improvements across all three translation directions. It is worth noting that while the baseline models showed a significant performance gap, this gap narrowed after applying GRPO, with SalamandraTA7B-Instruct even outperforming TowerInstruct-7B-v.02 on both English to German and English to Russian. These differences may be explained by the different behavior of the COMET-KIWI score during training. The improvement observed only in SalamandraTA7B-Instruct may have contributed to its final translation performance.

4.5 Discussion

To test our hypothesis that the translation-faithfulness reward functions as a regularization term, we trained SalamandraTA7B-Instruct using the same training configuration but optimizing only for the terminology adherence reward. Figure 3 illustrates the evolution of this reward in comparison to COMET-KIWI throughout training. While terminology adherence improves at a rate comparable to that observed in Figure 2, translation faithfulness declines rapidly after the initial training steps. This behavior suggests that, in the absence of a faithfulness reward, the model tends to produce translations that diverge semantically from the source sentence, resulting in degraded translation quality and increased hallucinations.

[2]We use the `fuzzywuzzy` Python library for fuzzy string matching.

[3]`Signature: nrefs:1– case:mixed– eff:no– tok:13a– smooth:exp–version:2.3.1`

[4]`Signature: nrefs:1– case:mixed– eff:yes– nc:6– nw:0–version:2.3.1`

[5]https://huggingface.co/Unbabel/wmt22-cometkiwi-da

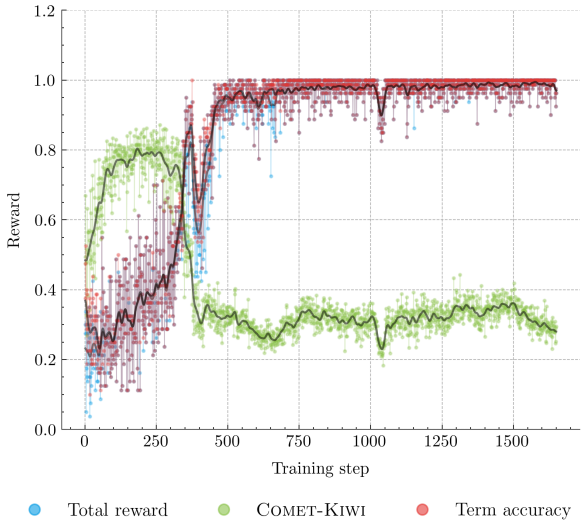


Figure 3: Rewards evolution during training when only the terminology adherence reward is optimized.

5 Conclusions

This paper presents the BSC team's submission to the shared task of terminology-aware Machine Translation. Our results show that GRPO training using only monolingual data can effectively adapt an LLM for this task, producing translations that include the correct terminology in almost all cases. Analysis of the training rewards shows the importance of including a quality-estimation term to regularize training and ensure strong translation performance. Studying the impact of the choice of languages and its relationship with performance remains an avenue for future work.

Limitations

All experiments in this work focus on high-resource languages (English, Spanish, German, and Russian). Quality-estimation metrics can be more consistent for these languages than for other low-resource counterparts. It is worth noting that some extremely low-resource languages may not be supported by any existing quality-estimation model. We leave it for future work to explore the robustness of the model across language families and low-resource scenarios.

Ethical Statement

This work focuses on the term accuracy and overall translation quality of the adapted models. The impact of this adaptation on possible biases, such as gender bias, produced by the system is outside the scope of this study.

All models and datasets used in these experiments are based on publicly available resources, and no direct causes of bias were observed.

Acknowledgements

This work has been promoted and financed by the Generalitat de Catalunya through the Aina Project.

This work has been supported by the Spanish project PID2021-123988OB-C33 funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is partially supported by MLLM4TRA (PID2024-158157OB-C32) funded by MCIN/AEI/10.13039/501100011033/FEDER, UE.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública - Funded by EU – NextGenerationEU within the framework of the ILENIA Project with reference 2022/TL22/00215337.

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023, Singapore, December 6-7, 2023*, pages 890–896. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhaopeng Feng, Shaosheng Cao, Jiahan Ren, Jiayuan Su, Ruizhe Chen, Yan Zhang, Zhe Xu, Yao Hu, Jian Wu, and Zuozhu Liu. 2025. MT-R1-Zero: Advancing LLM-based machine translation via R1-Zero-like reinforcement learning. *arXiv preprint arXiv:2504.10160*.

Javier Garcia Gilabert, Xixian Liao, Severino Da Dalt, Ella Bohman, Audrey Mash, Francesca De Luca Fornaciari, Irene Baucells, Joan Llop, Miguel Claramunt Argote, Carlos Escolano, and Maite Melero. 2025. From SALAMANDRA to SALAMANDRATA: BSC submission for WMT25 general machine translation shared task. *arXiv preprint arXiv:2508.12774*.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, and 5 others. 2025. Salamandra technical report. *arXiv preprint arXiv:2502.08489*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Miami, Florida, USA. Association for Computational Linguistics.

Zheng Li, Mao Zheng, Mingyang Song, and Wenjie Yang. 2025. TAT-R1: Terminology-aware translation with reinforcement learning and word alignment. *arXiv preprint arXiv:2505.21172*.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv: 2409.19256*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and*

Evaluation (LREC'12), pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jiawei Zheng, Hanghai Hong, Feiyan Liu, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *arXiv preprint arXiv:2402.15061*.

A Template

This section presents the template used to prepare instructions for training (Figure 4). We used only one single template. Placeholders:

- { Source Sentence }: source sentence
- { Glossary }: glossary of terms
- { Source Language }: source language name
- { Target Language }: target language name

Template used for training

```
Translate the following text from {Source Language} into {Target Language}.
Please ensure the following terminology is used:
{Glossary}.
{Source Language}: {Source Sentence}
{Target Language}:
```

Figure 4: Example of a template used to construct terminology instructions.

It Takes Two: A Dual Stage Approach for Terminology-Aware Translation

Akshat Singh Jaswal
PES University
sja.akshat@gmail.com

Abstract

This paper introduces DuTerm, a novel two-stage architecture for terminology-constrained machine translation. Our system combines a terminology-aware NMT model, adapted via fine-tuning on large-scale synthetic data, with a prompt-based LLM for post-editing. The LLM stage refines NMT output and enforces terminology adherence. We evaluate DuTerm on English-to-German, English-to-Spanish, and English-to-Russian for the WMT 2025 Terminology Shared Task. We demonstrate that flexible, context-driven terminology handling by the LLM consistently yields higher quality translations than strict constraint enforcement. Our results highlight a critical trade-off, revealing that an LLM's intrinsic knowledge often provides a stronger basis for high-quality translation than rigid, externally imposed constraints.

1 Introduction

The accurate and consistent translation of domain-specific terminology is a challenge in the field of Machine Translation and is of importance in domains such as law, medicine, and engineering, where precision is critical (Naveen and Trojovský, 2024). While modern Neural Machine Translation systems based on architectures like the Transformer have achieved remarkable fluency and quality on general text, their performance in terminology-constrained texts remains a critical area for improvement (Vaswani et al., 2023; Bahdanau et al., 2016; Johnson et al., 2017). This issue is particularly relevant given findings of recent WMT shared tasks, which have consistently highlighted the need for systems that can effectively handle domain-specific vocabulary (Post, 2018). The WMT 2025 Terminology Shared Task (Semenov et al., 2025) provides a focused platform to evaluate MT systems ability to handle domain-specific terminology under controlled conditions across multiple language pairs: English to German, English to Spanish, and English to Russian.

Previous research into terminology-constrained MT can be broadly categorized into two main approaches: inference-time methods and training-time methods. Inference-time approaches incorporate terminology constraints directly into the decoding process, often through techniques like constrained beam search or by re-ranking n-best lists of candidate translations (Zhang et al., 2023). While these methods are highly effective at enforcing constraints, they can be computationally expensive and may compromise the overall fluency and grammatical correctness of the output by forcing the model to generate awkward or unnatural phrases. Recent work has explored ways to make these methods more efficient, but the trade-off between enforcing terminology constraints and fluency remains a key consideration (Moslem et al., 2023).

Alternatively, training-time methods aim to teach models how to handle terminology constraints by integrating the terminology information into the training data itself. This is commonly done through the use of special tags that surround the terms to be translated (Dinu et al., 2019). This approach allows the model to learn how to produce more natural and grammatically correct output, but it provides no guarantee that all constraints will be respected during inference (Susanto et al., 2020).

We present DuTerm, a two-stage architecture that addresses these limitations by combining the strengths of both training-time and inference-time methodologies. We recognize that terminology-constrained translation is not merely a lexical substitution problem but requires a deeper understanding of linguistic context, especially when dealing with the complex morphology of languages like German and Russian. Our system is specifically designed to tackle the multifaceted evaluation framework of the WMT 2025 shared task.

Proceedings of the Tenth Conference on Machine Translation, pages 1344–1350

2 Method

2.1 Terminology-Aware Neural Machine Translation

Overview We develop a terminology aware MT model via large-scale, tagged synthetic data and targeted fine-tuning. The pipeline: extract and analyze terminology, generate tagged, context-rich parallel data (single-term and multi-term) and standardize tags and ensure annotation consistency. We also quality-filter with COMET_{QE} (Rei et al., 2022) and deduplicate , this finally adapts a multilingual NMT model with parameter efficient fine-tuning .

Terminology Extraction and Analysis We parse the WMT 2025 dev files for English→German/Spanish/Russian to build bilingual terminology dictionaries. The dictionaries typically exceed 1,000 unique pairs per direction. We track terms and occurrences using `repetition_ids`. We also use the LLM to generate more terms similar to the terms provided in the dictionaries.

Synthetic Data Generation We use GPT-4o (OpenAI, 2024) to create parallel sentences that naturally embed required terms and explicitly insert boundary tags (`[TERM]...[/TERM]`) on both source and target. There are two modes we use to generate these parallel sentences

Single-term mode: generates sentence pairs containing exactly one term instance per sentence.

Multi-term mode: randomly selects 2–3 term pairs to appear together, teaching co-occurrence handling and disambiguation.

We employ temperature sampling (0.3–0.7), concurrent generation, and strict parsing to yield well-formed bilingual pairs.

Tag Standardization and Quality Filtering A re-tagging pass enforces consistent annotation, longest-first matching prevents partial shadowing, case-insensitive detection with original case preservation, and inverse mapping ensures symmetric target-side tagging. Each pair is scored by COMET_{QE}. We deduplicate on the source side and keep only high-confidence items using a conservative threshold (0.85-0.9) depending on the language, typically retaining 60–70% of outputs, yielding ∼10k–15k pairs per language direction.

Multilingual Model Adaptation For the foundation translation model, we select NLLB-200 3.3B, a state-of-the-art multilingual neural machine translation model with demonstrated strong performance across our target languages (Team et al., 2022). This model provides robust baseline capabilities while supporting the specialized terminology handling adaptations we require.

We extend the model's vocabulary with terminology markup tokens to ensure atomic treatment of terminology annotations. This prevents subword tokenization from fragmenting our special markup, ensuring that terminology boundaries are consistently preserved during training and inference.

The training process employs several optimization strategies designed for stable, effective adaptation. The process also combines filtered datasets from all three target languages, creating unified multilingual adaptation that benefits from cross-lingual transfer.

2.2 LLM-Based Post-Editing

Overview An LLM refines the NMT output given the source sentence and required term pairs, enforcing strict terminology adherence while improving fluency and morphology.(Raunak et al., 2023)

Post-Editing Procedure We use prompts that present the source, translation, and provide explicit source to target term mappings. The LLM is instructed to preserve meaning, apply the exact target terms, maintain tags where required, and improve readability without paraphrasing away constraints. The LLM we choose to use is GPT-4o (OpenAI, 2024) due to its combination of high translation quality and relatively lower price.

Terminology-Aware Processing *Dynamic resolution*: per-input selection of proper/random/no-term constraints from reference terminology databases with whitespace-normalized matching. *Mode-adaptive behavior*: when constraints exist, the LLM must enforce them; otherwise it performs quality-only edits while being sensitive to technical terms. *Constraint satisfaction*: explicit mappings and formatting rules are included in the prompt; outputs must preserve required terminology and markup.

Quality Assurance and Robustness We run the LLM at low temperatures (0.3) for deterministic edits. Each hypothesis is validated for format, tag integrity, and constraint satisfaction before acceptance with a pre-existing parser. We verify filename

schemas, presence of all terminology modes per language pair, and JSONL structure. We assess quality with COMET_{QE} (after tag stripping) and compute terminology preservation via exact-match checks on required terms. This ensures reliability of final outputs.

3 Results

We evaluate the system using three complementary metrics used by the WMT organizers: BLEU for overall translation adequacy, chrF2++ for character-level fluency and robustness, and terminology success rates (proper and random) to directly measure constraint satisfaction (Papineni et al., 2002; Popović, 2015). Results are reported for English→German (DE), English→Spanish (ES), and English→Russian (RU) across three terminology strategies: *noterm*, *proper*, and *random*.

Table 1 summarizes the findings. Several clear patterns emerge:

1. **Strict terminology enforcement (*proper*)** achieves the highest BLEU and chrF2++ across all languages (48.06 for DE, 58.51 for ES, 35.80 for RU), indicating improved lexical precision and sentence-level quality when constraints are respected. It also yields near-perfect proper terminology success rates ($\geq$0.97).

2. **Unconstrained translation (*noterm*)** consistently underperforms, producing the lowest BLEU and chrF2++ values across languages (e.g., 38.24 BLEU in DE and 27.88 in RU). While fluency remains reasonable, failure to enforce constraints leads to poor terminology precision.

3. **Random terminology enforcement** produces intermediate BLEU/chrF2++ but near-perfect random-term success rates ($\sim$0.98). This highlights that while the model can force arbitrary terminology, doing so compromises contextual appropriateness.

4. **Language-specific trends** align with expectations: Spanish shows the highest overall scores, reflecting its structural similarity to English. Russian shows the widest gap between *proper* and *noterm*, emphasizing the difficulty of morphology-rich languages for terminology control.

Overall, these results demonstrate that while strict enforcement maximizes terminology accuracy and boosts surface-level quality metrics, it can occasionally reduce flexibility. In contrast, unconstrained approaches produce more natural translations but risk terminology inconsistency.

4 Conclusion

This paper presents our approach to the terminology shared task, focusing on English to German, Spanish, and Russian translation directions. Our system leverages LLMs to improve existing translations with varying terminology handling strategies. Our results demonstrate that allowing the LLM to flexibly handle terminology often yields higher translation quality than strict terminology enforcement. These findings highlight the potential of prompt-based LLM systems for technical and business translation tasks, and provide insights into effective strategies for terminology management in neural translation workflows. The intution behind why this approach works so well is that NMTs often excel at strict word-level translations however they can struggle with context-dependent nuances. Our approach leverages post-editing with a LLM on top of the initial NMT outputs. By starting from a reliable NMT translation, the post-editing model receives a structured, partially correct target sentence, which allows it to focus on higher-level improvements resolving ambiguities, adjusting word order, and refining context. This guided refinement is often more effective than translating directly from scratch with an LLM or NMT which must simultaneously handle term accuracy and contextual fluency.

For future work, exploring adaptive learning mechanisms that integrate terminology dynamically, rather than relying on static prompts, could enhance robustness across domains and languages. End-to-end or memory-augmented architectures that maintain consistency across sentences and documents hold promise for more coherent outputs. Expanding evaluations to other language models and diverse, domain-specific corpora would help validate the approach's generalizability and reveal domain-dependent challenges. Incorporating hybrid strategies, such as combining prompt guidance with fine-tuning or reinforcement learning, and enabling user-driven interaction for terminology control could further improve usability and accuracy. Together, these directions offer a pathway toward more flexible, context-aware, and widely applicable terminology-aware translation systems.

Limitations

Our approach, based on a prompt-driven framework, faces several limitations. It depends heavily on carefully crafted prompts, which may not gen-

Lang	Type	BLEU	chrF2++	Prop. SR	Rand. SR
DE	noterm	38.24	62.61	0.43	0.69
	proper	48.06	70.74	0.98	0.73
	random	43.77	67.22	0.48	0.99
ES	noterm	45.98	67.05	0.47	0.73
	proper	58.51	76.08	0.99	0.78
	random	53.28	72.05	0.49	0.98
RU	noterm	27.88	55.29	0.39	0.69
	proper	35.80	63.57	0.98	0.72
	random	32.25	59.85	0.42	0.99

Table 1: Evaluation results for English→German (DE), English→Spanish (ES), and English→Russian (RU) across three terminology handling strategies. Metrics include BLEU, chrF2++, and terminology success rates (proper and random).

eralize well across domains, languages, or model architectures. The sequential processing of terminology matching and translation refinement limits the system's ability to adaptively enforce terminology constraints. Furthermore, operating at the sentence level overlooks opportunities for document-level consistency and context-aware terminology usage, which are crucial in practical translation tasks. Our evaluation, conducted solely on GPT-4o (OpenAI, 2024), restricts the generalizability of findings, and focusing on technical and business domains may not capture challenges present in specialized fields like medical or legal translation. Additionally, while COMET_{QE} ,BLEU ,chrF++ provide automated scalability, it may not fully reflect terminological precision and contextual appropriateness, suggesting the need for complementary evaluation methods that include human judgment.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3893–3898, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Palanichamy Naveen and Pavel Trojovský. 2024. Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(10):110878.

OpenAI. 2024. Gpt-4o system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi,

United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada. Association for Computational Linguistics.

A Prompts

We include below the full prompts used in our experiments for reproducibility.

A.1 Single-Term Prompt

```
Generate {n} professional, domain-specific English-({target_lang}) bilingual sentence pairs for
    terminology translation.
The term pair to use is: {source_term}\(EN) : \"{target_term}\ ({target_lang})
Requirements:
- Each sentence pair must be natural, fluent, and contextually appropriate for IT or financial
    domains.
- Include the term exactly once per sentence.
- Wrap the term with [TERM] and [/TERM] in both the English and ({target_lang}) sentences.
- Ensure accurate translation and alignment of the term.
Format:
EN: [sentence]
{target_lang}: [sentence]
Output exactly {n} such pairs.
```

Listing 1: Prompt template for generating bilingual sentence pairs with a single terminology constraint.

A.2 Multi-Term Prompt

```
Generate {n} professional, domain-specific English-({target_lang}) bilingual sentence pairs for
    terminology translation.
Use ALL of the following term pairs in each sentence pair:\n{terms_str}
Requirements:
- Each sentence pair must be natural, fluent, and contextually appropriate for IT or financial
    domains.\n"
- Include each term exactly once per sentence.
- Wrap each term with [TERM] and [/TERM] in both the English and ({target_lang}) sentences.\n"
- Ensure accurate translation and alignment of the terms.
Format:
EN: [sentence]
{target_lang}: [sentence]
Output exactly {n} such pairs.
```

Listing 2: Prompt template for generating bilingual sentence pairs with multiple terminology constraints.

A.3 Post-Editing with Terminology

```
As an expert English-{target_lang} translator specializing in technical and business documentation,
    improve this {target_lang} translation.

SOURCE (English): {source}

CURRENT TRANSLATION ({target_lang}): {translation}

REQUIRED TERMINOLOGY (English: {target_lang}): {term_str}

YOUR TASK:
1. Ensure all required terminology is correctly used
2. Maintain the same meaning as the source text
3. Ensure natural, fluent {target_lang} that sounds like native content
4. Preserve formatting, numbers, and special characters
5. Match the tone and register of the original text

Return ONLY the improved {target_lang} translation with no explanations, notes, or other text.
```

Listing 3: Prompt for post-editing with explicit terminology mappings.

A.4 Post-Editing without Terminology

```
As an expert English-{target_lang} translator specializing in technical and business documentation,
    improve this {target_lang} translation.

SOURCE (English): {source}

CURRENT TRANSLATION ({target_lang}): {translation}

Note: There may be important terminology in the source text that should be translated precisely and
    consistently. Please ensure any technical or business terms are rendered correctly in {
    target_lang}.

YOUR TASK:
1. Enhance the translation for fluency and accuracy
2. Maintain the same meaning as the source text
3. Ensure natural, fluent {target_lang} that sounds like native content
4. Preserve formatting, numbers, and special characters
5. Match the tone and register of the original text

Return ONLY the improved {target_lang} translation with no explanations, notes, or other text.
```

Listing 4: Prompt for post-editing without explicit terminology guidance.

Author Index